Statistics Canada Statistique Canada

SURVEY METHODOLOGY

DECEMBER 1975

VOLUME 1 – NUMBER 2

A JOURNAL PRODUCED BY

HOUSEHOLD SURVEYS DEVELOPMENT DIVISION

STATISTICAL SERVICES FIELD

STATISTICS CANADA

SURVEY METHODOLOGY

December 1975 Volume 1 Number 2

A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.

CONTENTS

SURVEY METHODOLOGY

December 1975 Volume 1 Number 2

A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed, however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department Copies of papers in either Official Language will be made available upon request.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, 10th Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario - K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

CONTROLLED RANDOM ROUNDING

I.P. Fellegi
Assistant Chief Statistician, Statistical Services Field

Random rounding is a technique to ensure confidentiality of aggregate statistics. By randomly rounding all the components of a total, independently, together with the random rounding of the total itself, substantial discrepancies may arise when aggregating the published data. This paper presents a procedure which avoids substantial discrepancies while still protecting the concept of confidentiality.

1. INTRODUCTION

Random rounding is a technique to prevent statistical disclosure, both direct and residual. It consists of rounding published (or otherwise released) statistical aggregates to a multiple of some chosen base number -- but carrying out the rounding through a random mechanism which ensures that each randomly rounded published aggregate has as its expected value the corresponding unrounded (and, of course, unpublished) aggregate. This ensures that the rounding process is unbiased. For a more detailed description of the technique, the reader is referred to [2] and [3].

The particular problem addressed in this note can be summarized as follows. Given that each of a number of statistical aggregates has to be random rounded, can this be done in such a way that the sum of the individually random rounded numbers is equivalent to the random rounding of the sum of the unrounded numbers, i.e. if e_i $(i=1, 2, \ldots, n)$ are unrounded numbers, and e_i^* are the corresponding random rounded numbers, can we carry out the random rounding in such a way that

$$\sum_{i=1}^{n} e_i^*$$

is equivalent to (i.e. has the same distribution as)

$$\left(\sum_{i=1}^{n} e_i\right)^*.$$

The question so stated grew out of a very concrete problem. Several countries have adopted the practice of releasing so-called summary tapes after their decennial or quinquennial population censuses. These tapes contain tabulations (aggregates) at the level of very small geographic areas, usually corresponding to the work assignment of one census enumerator. These small area data are used by research personnel as "building blocks" to aggregate data for their respective areas of interest. At least two countries, the United Kingdom [4] and Canada, have adopted the practice of introducing a small random disturbance into these small area level aggregates in order to safeguard against statistical disclosure, and the Bureau of the Census is at least contemplating a similar procedure for the 1980 Census [1].

Even though the level of such random errors is small, when the random rounded numbers are aggregated, their variances aggregate also. When several small area tabulations are aggregated in order to obtain a tabulation for a large area, say a municipality, the variance may become quite large (although, of course, the relative variance declines). So when users compare their own tabulations prepared from the summary tapes for, say, a municipality, with the corresponding tables actually published at the level of a municipality, substantial discrepancies may be observed. The reason is that the published municipality-level tabulations were random rounded directly, while the tabulations prepared by users from the summary tapes were random rounded at the level of the component small area level.

The following procedure ensures that when the small-area tabulations are random rounded, the cumulative impact of such errors is controlled at the level of some predefined higher level geographical areas. Of course, for other than the predefined larger areas the variance due to random rounding is probably unaffected.

An attempt to contain the cumulative impact of random errors is given in [4], but only for a situation where the amount of random error is +1, -1 or 0.

2. THE CONSTRAINTS

If the base of random rounding is an integer b (be was equal to 5 in the 1971 Census in Canada), suppose that a table entry is e. We compute the residual r of e after division by b:

$$e = k \times b + r \qquad 0 \leq r < b$$

It is this residual which is "rounded" at random to either 0 or b. Let the probability of rounding up to b be p, the probability of rounding down to zero being (1-p). The randomly rounded e, e* can be written as

$$e^* = k \times b + r^*$$

where r* = b with probability p and it is equal to 0 with probability (1-p). The expected value of e* can be written as

$$E(e^*) = k \times b + [p \times b + (1-p) \times 0]$$

If we want e* to be unbiased, we must set

$$E(e^*) = e$$

i.e.

$$p \times b = r \quad \text{or} \quad p = \frac{r}{b}$$

This is the first constraint we impose on a desirable random rounding procedure. The argument above also shows that if e* is to be an unbiased estimate of e, the only way e can be altered to become a multiple of b while changing it at random by an amount which is less than b in absolute value, is by a random rounding process with probabilities as shown above.

If we want to preserve the unbiasedness of random rounding, this constraint must, therefore, not be violated.

Next, suppose there are a series of n tabulation cells (each corresponding to one small area aggregate in a municipality) which are to be rounded. Denote these by e_i $(i = 1, 2, \ldots, n)$.

Let

$$e_i = k_i x b + r_i \qquad 0 \leqq r_i < b$$

and the randomly rounded corresponding value as

$$e_i^* = k_i x b + r_i^*$$

where $r_i^* = b$ with probability $p_i = r_i/b$ and is equal to 0 with probability $1 - p_i$.

Their sum, e is

$$e = \sum_{i=1}^{n} e_i$$

which can be written as

$$e = k \times b + r \qquad 0 \leqq r < b \tag{2.1}$$

and its rounded value is

$$e^* = k \times b + r^*$$

where r^* is equal to b with probability of r/b and is zero otherwise. Ideally, one would like to have

$$e^* = \sum_{i=1}^{n} e_i^*$$

in the sense that $\Sigma\, e_i^*$ and e^* assume the same values with the same probabilities. This is the second constraint we impose on a desirable random rounding procedure.

The procedure below satisfies both of these.

3. THE PROCEDURE

Consider the numbers r_i and cumulate them:

$$s_i = \sum_{j=1}^{i} r_j$$

$$S_n = \sum_{j=1}^{n} r_j$$

$$S_o = 0$$

Select a random integer between 1 and b, say R_1:

$$1 \leqq R_1 \leqq b$$

Consider S_1, S_2, S_3, ..., in order until

$$S_{i_1-1} < R_1 \leqq S_{i_1}$$

Next let

$$R_2 = R_1 + b$$

and select i_2 so that

$$S_{i_2-1} < R_2 \leqq S_{i_2}$$

Next let

$$R_3 = R_2 + b = R_1 + 2b$$

and select i_3 so that

$$S_{i_3-1} < R_3 \leq S_{i_3}$$

etc. Continue until the L-th step so defined that

$$R_L \leq S_n$$

but

$$R_{L+1} > S_n$$

Now round <u>up</u> the units so selected, <u>down</u> the others. In other words,

$$e_i^* = k_i x b + r_i^*$$

where

$$r_i^* = b \qquad \text{if} \qquad i = i_1, i_2, i_3, \ldots$$

$$= 0 \qquad \text{otherwise.}$$

The procedure is illustrated in Table 1.

4. PROOF THAT THE PROCEDURE SO DEFINED SATISFIES THE CONSTRAINTS

It is easy to verify, using arguments which are standard in selecting with probabilities proportional to a measure of size, that the probability

$$P(r_i^* = b) = r_i/b$$

so that the first constraint is satisfied.

As far as the second constraint is concerned, the following simple argument shows that it, too, is satisfied.

Since from (2.1)

$$e = \sum_{i=1}^{n} e_i = kb + r$$

and also

$$e = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (k_i\, b + r_i) = b \sum_{i=1}^{n} k_i + \sum_{i=1}^{n} r_i$$

$$= b \sum_{i=1}^{n} k_i + S_n \tag{4.1}$$

it follows that the integer remainder of S_n, when divided by b, must also be r. So we must have, for some integer m,

$$S_n = mb + r \qquad 0 \leqq r < b \tag{4.2}$$

So from (4.1) we obtain

$$e = \sum_{i=1}^{n} e_i = b\,(\sum_{i=1}^{n} k_i + m) + r$$

i.e.

$$k = \sum_{i=1}^{n} k_i + m$$

It immediately follows from (4.2) that the number of steps, L, required to complete the procedure is related to m, r and R_1 as follows:

$$\text{Prob}\ (L = m + 1) = \text{Prob}\ (1 \leqq R_1 \leqq r) = \frac{r}{b}$$

$$\text{Prob}\ (L = m) = \text{Prob}\ (r < R_1 \leqq b) = 1 - \frac{r}{b}$$

Since

$$\sum_{i=1}^{n} e_i^* = b \sum_{i=1}^{n} k_i + \sum_{i=1}^{n} r_i^* = b \sum_{i=1}^{n} k_i + Lb$$

we have

$$\sum_{i=1}^{n} e_i^* = \begin{cases} b \sum_{i=1}^{n} k_i + (m+1)\, b & \text{with probability } \frac{r}{b} \\ b \sum_{i=1}^{n} k_i + mb & \text{with probability } 1 - \frac{r}{b} \end{cases}$$

$$= \begin{cases} kb + b & \text{with probability } \frac{r}{b} \\ kb & \text{with probability } 1 - \frac{r}{b} \end{cases} \tag{4.3}$$

Also,

$$\left(\sum_{i=1}^{n} e_i\right)^* = kb + r^*$$

where

$$P(r^* = b) = \frac{r}{b}$$

$$P(r^* = 0) = 1 - \frac{r}{b}$$

so that

$$\left(\sum_{i=1}^{n} e_i\right)^* = \begin{cases} kb + b & \text{with probability } \frac{r}{b} \\ kb & \text{with probability } 1 - \frac{r}{b} \end{cases} \tag{4.4}$$

Comparing (4.3) and (4.4) we obtain immediately that the random variables $\Sigma\, e_i^*$ and $(\Sigma\, e_i)^*$ have the same distribution.

Thus the net effect of the procedure on a predefined aggregate of randomly rounded individual numbers is equivalent to the random rounding of the aggregate itself.

It can also be shown quite readily that the same argument holds for the sum of any consecutive numbers $e_t, e_{t+1}, \ldots, e_{t+5}$. Thus controlled random rounding results in a desirable reduction of rounding variance not only for a predefined aggregate, but also for any user-defined area consisting of the union of consecutive "building block" areas.

Table 1: Example of Controlled Random Rounding

											Total
Unrounded E.A. total (e_i)	12	23	34	3	49	23	50	17	8	13	232
Unroundable "base" ($k_i b$)	10	20	30	0	45	20	50	15	5	10	205
Residual (r_i)	2	3	4	3	4	3	0	2	3	3	27
Cumulative Residual (S_i)	2	5	9	12	16	19	19	21	24	27	
$R_1 = 1$	*		*	*	*			*		*	
Rounded Residual (r_i^*)	5	0	5	5	5	0	0	5	0	5	30
Rounded E.A. total (e_i^*)	15	20	35	5	50	20	50	20	5	15	235
$R_1 = 2$	*		*	*		*			*	*	
Rounded Residual (r_i^*)	5	0	5	5	0	5	0	0	5	5	30
Rounded E.A. total (e_i^*)	15	20	35	5	45	25	50	15	10	15	235
$R_1 = 3$		*	*		*	*			*		
Rounded Residual (r_i^*)	0	5	5	0	5	5	0	0	5	0	25
Rounded E.A. total (e_i^*)	10	25	35	0	50	25	50	15	10	10	230
$R_1 = 4$		*	*		*	*			*		
Rounded Residual (r_i^*)	0	5	5	0	5	5	0	0	5	0	25
Rounded E.A. total (e_i^*)	10	25	35	0	50	25	50	15	10	10	230
$R_1 = 5$		*		*	*			*		*	
Rounded Residual (r_i^*)	0	5	0	5	5	0	0	5	0	5	25
Rounded E.A. total (e_i^*)	10	25	30	5	50	20	50	20	5	15	230
No. of times Rounded up	2	3	4	3	4	3	0	2	3	3	2
No. of times Rounded down	3	2	1	2	1	2	5	3	2	2	3

RESUME

L'arrondissement aléatoire est une technique qui vise à assurer la confidentialité des agrégats ou groupes de statistiques. En appliquant cette technique à tous les éléments d'un total, d'une part, et au total lui-même, d'autre part, des divergences importantes peuvent se produire au moment de regrouper les données publiées. La méthode décrite dans ce document permet d'éviter ces divergences tout en assurant la confidentialité des données.

REFERENCES

[1] Barabba, V.P. and Kaplan, D.C., "U.S. Census Bureau statistical techniques to prevent disclosure -- the right to privacy vs. the need to know". Paper presented at the 2nd meeting of the international Association of Survey Statisticians, Warsaw (1975).

[2] Fellegi, I.P. and Phillips, J.L., "Statisticians Confidentiality: Some Theory and Applications to Data Dissemination". Annals of Economic and Social Measurement, pp. 399-409, 1974.

[3] Nargundkar, M.S. and Saveland, W., "Random Rounding: A Means of Preventing Disclosure of Information about Individual Respondents in Aggregate Data". Proceedings of the Social Statistics Section of ASA, 1972.

[4] Newman, D., "Rounding and Error Injection of Preserving Confidentiality of Census Data". Paper presented at the 2nd meeting of the International Association of Survey Statisticians, Warsaw (1975).

ON A RATIO ESTIMATE WITH POST-STRATIFIED WEIGHTING

G.B. Gray and P.D. Ghangurde
Household Surveys Development Division

A ratio estimate based on an auxiliary variable is considered for the case when the sample is post-stratified using information on another auxiliary variable. The variance of the ratio estimate is derived by the method of linearization [3,4]. An application to subprovincial estimation in the Canadian Labour Force Survey is discussed.

1. INTRODUCTION

Consider a population which is stratified into L strata. Let y and x be respectively the variable of interest and auxiliary variable. Let samples drawn independently from strata be post-stratified into k strata by using information on another auxiliary variable, z, obtained for the samples. It is known that ratio y/x is more homogeneous within post-strata defined by z than within strata. The ratio estimate for each stratum post-stratum cell can be weighted by x-totals to obtain an estimate of y-total for these cells. However, the weights i.e., x-totals are not available at the level of stratum-post-stratum cells but at the higher level of groups of strata for each post-stratum. These weights can be used in ratio estimation for obtaining estimates of y-totals for a stratum or a group of strata for any multistage design within strata.

The situation occurs in surveys of human populations in which geographic areas are used as strata. The samples drawn from strata can be post-stratified by using information on characteristics like age, sex, ethnic origin etc. as it is known that ratio y/x, usually proportion of population with a socio-economic characteristic, is more homogeneous within these subclasses than within the original strata. Since it is difficult to select and control the sample within these subclasses post-stratification has to be resorted to. The population within these subclasses may be known at the level of province, state, etc. rather than at the lower level of economic regions, enumeration areas, etc. for which estimates of characteristic totals are needed. Many times these population figures

within subclasses are estimates based on a recent census and are likely to be more reliable than estimates of population obtained from sample survey.

2. NOTATIONS AND ESTIMATE

Let

$\hat{Y}_{ij}$ = estimate of y-total in (i, j)th cell,
$i = 1, 2, \ldots, L,$
$j = 1, 2, \ldots, k,$

$\hat{X}_{ij}$ = estimate of x-total in (i, j)th cell
$i = 1, 2, \ldots, L$
$j = 1, 2, \ldots, k,$

a and b be two sets of strata $b \subseteq a$,

X_{aj} = x-total in group of cells $\sum_{i \varepsilon a}$ (i, j).

The sample design within strata could be any multistage design. The estimates $\hat{Y}_{ij}$ and $\hat{X}_{ij}$ are obtained for cells (i, j) which can be considered as domains.

The ratio estimate of y-total in set b is obtained by weighting separate ratio estimates over post-strata and is given by

$$\hat{Y}_b = \sum_{j=1}^{k} \left[\frac{\sum_{i \varepsilon b} \hat{Y}_{ij}}{\sum_{i \varepsilon a} \hat{X}_{ij}} \right] \cdot X_{aj} \qquad (2.1)$$

It may be noted that ratio estimates $\sum_{i \varepsilon b} \hat{Y}_{ij} / \sum_{i \varepsilon a} \hat{X}_{ij}$ are not the ratio estimates in the usual sense unless b = a. However, this type of ratio estimate can yield gains in precision if ratios are high enough. The weights X_{aj} at the level of 'a' can thus be used for ratio estimation at a lower level.

3. VARIANCE OF $\hat{Y}_b$

The procedure of linearization of the ratio for expressing variance was used by Keyfitz [1] to obtain variances for specific designs and later by Kish [2] for variances of indexes of complex samples. A generalization of the method using Taylor series approximation and for any design is given by Woodruff [4]. The ratio approximation of $\hat{Y}_b$ given below is the same as one which can be obtained by Taylor series. It is assumed that the sample within cells is large enough to justify the approximation.

Let
$$F_j = \left(\sum_{i\varepsilon b} \hat{Y}_{ij}\right)/\left(\sum_{i\varepsilon a} \hat{X}_{ij}\right) \tag{3.1}$$

be the ratio estimate for the jth post-stratum, $j = 1, 2, \ldots, k$.

The linearized form of F_j, denoted by G_j, is given by the ratio approximation as

$$G_j \doteq \left[\sum_{i\varepsilon b} \hat{Y}_{ij} - R_j \sum_{i\varepsilon a} \hat{X}_{ij}\right]/\left(\sum_{i\varepsilon a} \hat{X}_{ij}\right), \qquad j = 1, 2, \ldots, k. \tag{3.2}$$

where $R_j = \left(\sum_{i\varepsilon b} Y_{ij}\right)/\left(\sum_{i\varepsilon a} X_{ij}\right)$.

$$\begin{aligned}
\hat{Y}_b &\doteq \sum_{j=1}^{k} G_j X_{aj} \\
&= \sum_{j=1}^{k} \left[\sum_{i\varepsilon b} \hat{Y}_{ij} - R_j \sum_{i\varepsilon a} \hat{X}_{ij}\right] \\
&= \sum_{i\varepsilon b} \left[\sum_{j=1}^{k} (\hat{Y}_{ij} - R_j \hat{X}_{ij})\right] - \sum_{i\varepsilon c} \sum_{j=1}^{k} R_j \hat{X}_{ij},
\end{aligned} \tag{3.3}$$

where $c = a - b$. The change of order of summation in $\hat{Y}_b$ avoids the computation of variances and covariances for post-strata within strata as explained in Woodruff [4].

Since sampling is done independently within each stratum,

$$V(\hat{Y}_b) \doteq \sum_{i\varepsilon b} V[\sum_{j=1}^{k} (\hat{Y}_{ij} - R_j \hat{X}_{ij})] + \sum_{i\varepsilon c} V(\sum_{j=1}^{k} R_j \hat{X}_{ij}). \tag{3.4}$$

The second term in (3.4) appears due to weighting at a higher level than the level of estimation. An alternative to ratio estimate would be simple estimate of y-total appropriate for the design used and is given by $\hat{\hat{Y}}_b$ where

$$\hat{\hat{Y}}_b = \sum_{i\varepsilon b} \hat{Y}_i , \tag{3.5}$$

where $\hat{Y}_i$ = estimate of y-total in the ith stratum, and

$$V(\hat{\hat{Y}}_b) = \sum_{i\varepsilon b} V(\hat{Y}_i). \tag{3.6}$$

Hence $\hat{Y}_b$ is more efficient than $\hat{\hat{Y}}_b$ if

$$2 \sum_{i\varepsilon b} \mathrm{Cov} (\sum_{j=1}^{k} \hat{Y}_{ij}, \sum_{j=1}^{k} R_j \hat{X}_{ij}) > \sum_{i\varepsilon b} V(\sum_{j=1}^{k} R_j \hat{X}_{ij}) + \sum_{i\varepsilon c} V(\sum_{j=1}^{k} R_j \hat{X}_{ij}). \tag{3.7}$$

Since ratio estimation with post-stratified weighting takes advantage of homogeneity of ratios within post-strata the inequality (3.7) can be satisfied even when c is comparatively large if ratios R_j are large enough.

The variance expression in (3.4) can be simplified as

$$V(\hat{Y}_b) = \sum_{i\varepsilon a} V[\sum_{j=1}^{k} ({}^{b}\hat{Y}_{ij} - R_j \hat{X}_{ij})] \tag{3.8}$$

where

$${}^{b}\hat{Y}_{ij} = \begin{cases} \hat{Y}_{ij} & \text{if } i\varepsilon b \\ 0 & \text{if } i\varepsilon c \end{cases} \tag{3.9}$$

Thus $V(\hat{Y}_b)$ for any design within strata and any set b can be written as

$$V(\hat{Y}_b) = \sum_{i \varepsilon a} V(\hat{Y}_i'), \tag{3.10}$$

where
$$\hat{Y}_i' = \sum_{j=1}^{k} ({}^b\hat{Y}_{ij} - R_j \hat{X}_{ij}). \tag{3.11}$$

The formula (3.10) can be obtained from (3.6) by changing $\hat{Y}_i$ to $\hat{Y}_i'$ and b to a.

4. VARIANCE ESTIMATE

The variance estimate of $\hat{Y}_b$, $V(\hat{Y}_b)$, can be written down from variance estimate of $\hat{\hat{Y}}_b$ for stratified sampling.

Since R_j's are not known F_j's (as in 3.1) can be substituted as estimates of R_j's [4] and hence

$$V(\hat{Y}_b) = \sum_{i \varepsilon a} V(\hat{Y}_i'') \tag{4.1}$$

where

$$\hat{Y}_i'' = \sum_{j=1}^{k} ({}^b\hat{Y}_{ij} - F_j \hat{X}_{ij}) \tag{4.2}$$

Thus both variance and variance estimate can be written down by simple substitutions in the corresponding formulas for estimate $\hat{\hat{Y}}_b$, which does not use auxiliary information.

5. APPLICATION

The ratio estimate (2.1) is used for estimating totals for various labour force characteristics at subprovincial levels in the Labour Force Survey. The weights X_{aj} are estimates of population in various age-sex groups at provincial levels. These are based on a recent census and use demographic data like births, deaths, emigration, immigration and

population movements to arrive at projected population figures in various age-sex groups at provincial level. These estimates are available each month.

In the case of Labour Force Survey the assumption that $E(\sum_{i\varepsilon a} \hat{X}_{ij}) = X_{aj}$ made in the derivation of (3.3) may not hold due to possible coverage errors in the sampling frame used in selecting the area sample. A measure of these coverage errors in the frame is 'slippage'. If $E[\sum_{i\varepsilon a} \hat{X}_{ij}] = X^*_{aj}$ the 'slippage' at level a for jth age-sex group is defined as $(X_{aj} - X^*_{aj}).100/X_{aj}$. The weighting of the ratios within post-strata by population X_{aj} is expected to reduce the non sampling variance due to these coverage errors in the frame in addition to sampling variance reduction achieved due to ratio estimation.

The $V(\hat{Y}_b)$ given in (3.9) gets modified to

$$V(\hat{Y}_b) = \sum_{i\varepsilon a} V[\sum_{i=1}^{k} ({}^b\hat{Y}_{ij} - R_j \hat{X}_{ij}) \frac{X_{aj}}{X^*_{aj}}] \qquad (5.1)$$

There are analogous modifications in the definitions of $\hat{Y}'_i$ and $\hat{Y}''_i$ given in (3.11) and (4.2) and inequality (3.7).

Table 1 gives subweighted estimates, $\hat{\hat{Y}}_b$, final weighted estimates, $\hat{Y}_b$, and their % coefficient of variation (C.V.) for five important characteristics from January 75 survey in Ontario.

The characteristics are:

Unemployed, Employed, Employed-Agriculture, Employed-Non agriculture and In Labour Force. The areas considered are progressively increased from one region to the whole province which consists of ten regions.

It can be seen that the final weighted estimates are higher than the subweighted estimates by about 5% due to the correction for slippage in post-stratified weighting in final estimates. The C.V. for subweighted estimates is higher than that for final weighted estimates for "Employed",

"Employed Non agriculture" and "In Labour Force" which are the characteristics with high correlation with population count. There is a loss in efficiency by final weighting for "Unemployed" for small areas i.e. below regions 0 to 4 and very small gains for larger areas. In the case of "Employed Agriculture", which is the characteristic with lowest correlation with population count, there is in general a loss of efficiency due to final weighting. The gain in efficiency due to final weighting increases as the area considered is increased from single region to the whole of province.

6. ACKNOWLEDGMENT

The authors wish to thank M. Lawes for programming help and the referee for some helpful comments.

Table 1: Comparison of Subweighted and Final Weighted Estimates

Region	Characteristic	Subweighted		Final Weighted		Efficiency
		Estimate	C.V. %	Estimate	C.V. %	
0	Employed	378298	5.03	397390	4.85	1.04
	Unemployed	23551	12.23	24792	12.54	0.98
	Emp. Ag.	14829	24.24	15685	24.20	1.00
	Emp. Non Ag.	363468	5.19	381705	5.01	1.04
	In LF	401849	4.70	422182	4.54	1.04
0 to 1	Employed	498889	4.17	524467	4.02	1.04
	Unemployed	30641	11.04	32231	11.22	0.98
	Emp. Ag.	25889	16.06	27376	16.09	1.00
	Emp. Non Ag.	473000	4.44	497090	4.28	1.04
	In LF	529530	3.92	556698	3.79	1.03
0 to 2	Employed	1838265	2.69	1929436	1.93	1.39
	Unemployed	122817	5.89	129064	6.03	0.98
	Emp. Ag.	37588	11.55	39720	11.66	0.99
	Emp. Non Ag.	1800678	2.76	1889716	1.99	1.39
	In LF	1961083	2.51	2058501	1.77	1.42
0 to 3	Employed	2223012	2.33	2333670	1.59	1.47
	Unemployed	167871	4.83	176528	4.96	0.97
	Emp. Ag.	46747	10.51	49262	10.57	0.99
	Emp. Non Ag.	2176265	2.38	2284408	1.64	1.45
	In LF	2390883	2.17	2510198	1.47	1.48
0 to 4	Employed	2416490	2.18	2536681	1.47	1.48
	Unemployed	183971	4.59	193508	4.70	0.98
	Emp. Ag.	61820	11.45	65075	11.63	0.98
	Emp. Non Ag.	2354670	2.25	2471606	1.58	1.50
	In LF	2600461	2.04	2730189	1.35	1.51
0 to 5	Employed	2628910	2.17	2759682	1.20	1.81
	Unemployed	207525	4.60	218355	4.46	1.03
	Emp. Ag.	68682	10.63	72274	10.73	0.99
	Emp. Non Ag.	2560228	2.08	2687408	1.22	1.70
	In LF	2836434	1.93	2978037	1.08	1.79

Region	Characteristic	Subweighted		Final Weighted		Efficiency
		Estimate	C.V. %	Estimate	C.V. %	
0 to 6	Employed	2845985	2.06	2988047	1.11	1.86
	Unemployed	217806	4.61	229111	4.31	1.07
	Emp. Ag.	88632	6.90	93282	9.44	0.73
	Emp. Non Ag.	2757353	2.12	2894765	1.14	1.86
	In LF	3063790	2.00	3217158	1.01	1.98
0 to 7	Employed	2999218	2.06	3148947	0.79	2.61
	Unemployed	234445	4.61	246645	4.23	1.09
	Emp. Ag.	98159	6.90	103307	9.02	0.76
	Emp. Non Ag.	2901059	2.12	3045640	0.82	2.59
	In LF	3233663	2.00	3395592	0.66	3.03
0 to 8	Employed	3192817	1.95	3352861	0.71	2.75
	Unemployed	246797	4.46	259470	4.11	1.09
	Emp. Ag.	99514	8.79	104734	8.91	0.99
	Emp. Non Ag.	3093303	2.00	3248126	0.72	2.78
	In LF	3439614	1.90	3612331	0.56	3.39
0 to 9	Employed	3266860	1.92	3430734	0.67	2.87
	Unemployed	251331	4.39	264292	4.05	1.08
	Emp. Ag.	101111	8.76	105365	8.87	0.99
	Emp. Non Ag.	3166749	1.96	3325369	0.68	2.88
	In LF	3518191	1.86	3695026	0.52	3.58

RESUME

On établit une estimation du ratio en se fondant sur une variable auxiliaire lorsque l'échantillon est stratifié après coup au moyen des renseignements obtenus pour une autre variable auxiliaire. La variance de l'estimation du ratio se calcule par une mise sous forme linéaire [4,3]. On étudie le cas de l'estimation intraprovinciale dans l'enquête sur la population active.

REFERENCES

[1] Keyfitz, N. (1957), "Estimates of Sampling Variance when Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52, pp. 503-510.

[2] Kish, L. (1968), "Standard Errors of Indexes from Complex Samples", Journal of the American Statistical Association, 63, pp. 512-529.

[3] Tepping, B.J. (1968), "Variance Estimation in Complex Surveys", Proceedings of the Social Statistics section, ASA, pp. 11-18.

[4] Woodruff, R.S. (1971), "Simple Method for Approximating Variance of A Complicated Estimate", Journal of the American Statistical Association, pp. 411-414.

MEASUREMENT OF RESPONSE ERRORS IN CENSUSES AND SAMPLE SURVEYS

G.J. Brackstone, J.F. Gosselin, B.E. Garton
Census Survey Methods Division

Madow [1968] has proposed a two-phase sampling scheme under which response bias can be eliminated from sample surveys by obtaining 'true' values for a subsample of the original sample. Often in cases of Censuses or ongoing surveys, the subsample data are not used to correct the main survey estimates but to assess their reliability. The main purpose of this paper is to present methods by which reliability estimates can be obtained when true values can be determined for a subsample of units.

1. INTRODUCTION

A sample scheme was proposed by Madow [1965] under which the response bias could be eliminated from sample survey estimates by obtaining 'true observations' for a subsample of the original sample. This is achieved by using the estimate of bias from the subsample to correct the original estimate.

There are some instances, however, where the subsample data are obtained for evaluation purposes after the survey data has been published. In this case the objective is the measurement of the overall reliability of previously published survey data for the purpose of

a) informing the user of the data of its reliability (allowing him to make adjustments to it if he wishes), and,

b) informing the survey-taker of the sources of error in the survey so that improvements can be made in future surveys.

The purpose of this paper is therefore to present estimators of the reliability of Census and sample survey data when true values can be obtained for a sample or sub-sample of cases.

Two approaches to this problem are considered. In the first case, we consider the problem under the usual 'response variance - response bias' framework. This represents the main part of the paper. The second approach ignores the above model and attempts to measure only the net error in the particular survey observed.

The results in this paper apply to any sampling scheme for the original survey and to any sub-sampling scheme for true values, provided only that estimators of sampling variance are available for the sampling schemes used.

2. APPROACH I

2.1 Response Error Model

The response error model is based on the concept of independent repetition. We will first treat the Census case (i.e. 100% enumeration). Suppose, hypothetically, that many independent 'trials' of the Census could be made under the same general conditions. The Census 'estimate' for a given category of interest would then follow a certain frequency distribution. A particular Census figure may then be considered to be a random observation from a distribution of all possible Census estimates.

Let $\bar{X}(t)$ denote the Census estimate obtained at trial t and let $\bar{\mu}$ denote the corresponding true mean. The usual statistical parameter used to measure the reliability of an estimate in this situation is the Mean Square Error (MSE) of $X(\bar{t})$:

$$\text{MSE}\ (\bar{X}(t)) = \underset{t}{E}\ (\bar{X}(t) - \bar{\mu})^2$$

$$= \underset{t}{V}\ [\bar{X}(t)] + B^2$$

where B denotes the bias of $\bar{X}(t)$,

$$\text{i.e.} \quad B = \underset{t}{E}\,[\bar{X}(t)] - \bar{\mu}$$

$$= \bar{X} - \bar{\mu} \quad \text{where} \quad \bar{X} = \underset{t}{E}\,[\bar{X}(t)]$$

The response variance of $\bar{X}(t)$ is defined as

$$\underset{t}{V}\,(\bar{X}(t)) = \underset{t}{E}\,(\bar{X}(t) - \bar{X})^2$$

In general, therefore, bias results from errors that tend to occur in one direction rather than another. For example, if an error was present in the instructions accompanying the Census questionnaire, errors would tend to be systematic in one direction. Hence, bias measures the average net effect of all these possible factors.

On the other hand, response variance measures the random component of the error. For instance, in the case of an ambiguous question, a self-enumerated person may give different responses on different independent trials. These types of error depend on unknown factors that are impossible to control and may vary from trial to trial (e.g. the frame of mind of the respondent, the fatigue of the interviewer).

The above discussion applies to any characteristic obtained from a Census. However, it can easily be extended to cover characteristics obtained from a sample survey. In this case, let $\bar{x}_s(t)$ denote the estimate obtained from sample, s, at trial t. The MSE $[\bar{x}_s(t)]$ can then be expressed as

$$\text{MSE}\,[\bar{x}_s(t)] = E\,[\bar{x}_s(t) - \bar{\mu}]^2$$

where the expectation is taken over all possible trials and samples.

Letting $\bar{X} = E\,(\bar{x}_s(t)) = \underset{s}{E}\,\underset{t}{E}\,(\bar{x}_s(t)|s)$

and $\quad B = \bar{X} - \bar{\mu}$.

$$\text{MSE}\,[\bar{x}_s(t)] = E\,(\bar{x}_s(t) - \bar{X})^2 + B^2$$

$$= \underset{s}{E}\,\underset{t}{V}\,(\bar{x}_s(t)|s) + \underset{s}{V}\,\underset{t}{E}\,(\bar{x}_s(t)|s) + B^2$$

where the first term measures response variance, the second measures sampling variance, and the third measures bias.

2.2 Estimation

In the response error model described above three statistical parameters are defined, the mean square error, the response variance and the bias. Ideally we would like to obtain estimates of all three of these parameters in order to assess the level of both random and systematic errors and to obtain an overall measure of reliability.

Under the assumption that the true value of a sample characteristic is known for a random sub-sample of the survey sample, unbiased estimates of the MSE and bias are derived below. An unbiased estimator of the true proportion, $\bar{\mu}$, is also given following Madow [1965]. However, under this framework, it is not possible to estimate the total response variance.

Suppose true values are known for a random sub-sample, s', from s. Let $\bar{x}_{s'}(t)$ denote the unbiased estimate made from the sub-sample s' using the values observed on trial t, and let $\bar{\mu}_{s'}$ denote the corresponding estimate using the true values.

So $\qquad \underset{s'}{E}\,(\bar{x}_{s'}(t)|s,t) = \bar{x}_s(t)$

$$\underset{s}{E}\,\underset{s'}{E}\,(\bar{\mu}_s) = \bar{\mu}$$

Thus $\hat{B} = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ is an unbiased estimator of B.

Consider first the estimator $(\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2$. Letting E denote expectations over s', s and t (i.e. $E = \underset{t}{E}\,\underset{s}{E}\,\underset{s'}{E}$).

$$E(\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2 = V(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) + [E(\bar{x}_{s'}(t) - \bar{\mu}_{s'})]^2$$

$$= \underset{t}{V}\{\underset{ss'}{E}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\} + \underset{t}{E}\,\underset{ss'}{V}\{(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\} + B^2$$

Secondly, let $v_s(\bar{x}_s(t))$ be a variance estimator such that

$$\underset{s}{E}\,[v_s(\bar{x}_s(t))|t] = \underset{s}{V}\,[\bar{x}_s(t)|t]$$

i.e. an unbiased estimator of the sampling variance over the survey sampling scheme for a given trial t.

Thirdly, let $v_{ss'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ be a variance estimator such

$$\underset{ss'}{E}\,[v_{ss'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t] = \underset{ss'}{V}\,[(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t]$$

Now, if we define $\hat{MSE}\,(\bar{x}_s(t))$ by

$$\hat{MSE}(\bar{x}_s(t)) = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2 + v_s(\bar{x}_s(t)) - v_{ss'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) \tag{2.1}$$

then

$$E(\hat{MSE}\,(\bar{x}_s(t))) = \underset{t}{V}\,\underset{ss'}{E}\,\{(x_{s'}(t) - \bar{\mu}_{s'})|t\} + \underset{t}{E}\,\underset{ss'}{V}\,\{(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\}$$

$$+ B^2 + \underset{t}{E}\,\underset{s}{V}(\bar{x}_s(t)|t) - \underset{t}{E}\,\underset{ss'}{V}\,\{(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t\}$$

$$= \underset{t}{V}\,\underset{s}{E}\,(\bar{x}_s(t)|t) + \underset{t}{E}\,\underset{s}{V}\,(\bar{x}_s(t)|t) + B^2$$

$$= V_{t,s}(\bar{x}_s(t)) + B^2$$

$$= MSE\,(\bar{x}_s(t))$$

Thus $\hat{MSE}\,(\bar{x}_s(t))$ given by (2.1) is an unbiased estimator of $MSE\,(\bar{x}_s(t))$.

In the case where the original survey is a Census, the middle term in $\hat{MSE}\,(\bar{x}_s(t))$ disappears and we have

$$\hat{MSE}(\bar{X}(t)) = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})^2 - v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) \tag{2.2}$$

Given the unbiased estimator, $\hat{B}$, of the bias of $\bar{x}_s(t)$, an unbiased estimator of the true proportion, $\bar{\mu}$, is given by

$$\hat{\bar{\mu}} = \bar{x}_s(t) - \hat{B}$$

$$= \bar{x}_s(t) - (\bar{x}_{s'}(t) - \bar{\mu}_{s'}),$$

with $$V(\hat{\bar{\mu}}) = \underset{t}{E}\,\underset{s}{E}\,\underset{s'}{V}\,(\hat{\bar{\mu}}) + \underset{t}{E}\,\underset{s}{V}\,\underset{s'}{E}\,(\hat{\bar{\mu}}) + \underset{t}{V}\,\underset{s}{E}\,\underset{s'}{E}\,(\hat{\bar{\mu}})$$

$$= \underset{t}{E}\,\underset{s}{E}\,\underset{s'}{V}\,(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) + \underset{t}{E}\,\underset{s}{V}\,\underset{s'}{E}\,(\bar{\mu}_{s'})$$

since $\underset{s'}{E}\,(\hat{\bar{\mu}}|t,s) = \underset{s'}{E}\,(\bar{\mu}_{s'}|t,s)$

and $\underset{t}{V}\,\underset{s}{E}\,\underset{s'}{E}\,(\hat{\bar{\mu}}) = 0$

$$\therefore\ V(\hat{\bar{\mu}}) = \underset{t}{E}\,\underset{s}{E}\,\underset{s'}{V}\,(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) + \underset{t}{E}\,\underset{ss'}{V}\,(\bar{\mu}_{s'}) - \underset{t}{E}\,\underset{s}{E}\,\underset{s'}{V}\,(\bar{\mu}_{s'})$$

If $v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ is a variance estimator such that

$$E\{v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'})|t,s\} = V_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}),$$

if $v_{ss'}(\bar{\mu}_{s'})$ is a variance estimator such that

$$E[v_{ss'}(\bar{\mu}_{s'})|t] = V_{ss'}(\bar{\mu}_{s'})$$

and if $v_{s'}(\bar{\mu}_{s'})$ is a variance estimator such that

$$E[v_{s'}(\bar{\mu}_{s'})|s,t] = V_{s'}(\bar{\mu}_{s'})$$

then an unbiased estimator of $V(\hat{\bar{\mu}})$ is given by

$$\hat{V}(\hat{\bar{\mu}}) = v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) - v_{s'}(\bar{\mu}_{s'}) + v_{ss'}(\bar{\mu}_{s'})$$

In the case where the original survey is a Census, an unbiased estimator, $\hat{B}$, of the bias of $\bar{X}(t)$ is again given by

$$\hat{B} = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

and an unbiased estimator of the true proportion is given by

$$\hat{\bar{\mu}} = \bar{X}(t) - (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

In the expression for the variance of $\hat{\bar{\mu}}$ derived above for the sample survey case, s now becomes the total population. The second and third terms therefore cancel and we get

$$V(\hat{\bar{\mu}}) = E\ V\ (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

and therefore an unbiased estimator of $V(\hat{\bar{\mu}})$ is given by

$$V(\hat{\bar{\mu}}) = v_{s'}[\bar{x}_{s'}(t) - \bar{\mu}_{s'}]$$

2.3 Example

The purpose of this section is to describe how the previous estimators have been applied to a small study that was carried out in connection with the 1971 Census, and to present some numerical results.

When the 1971 Census data on type of dwelling were obtained it was suspected that certain categories, namely apartments and duplexes, had been grossly under-reported. As a result, a series of small scale studies were undertaken in an attempt to identify the sources of error.

One of these studies was carried out in the Ottawa region. One of its objectives was to compare the respondents' answers to the type of dwelling question in the 1971 Census with the 'true' type of dwelling as determined by visual observation by an expert. This comparison was carried out on all households in twelve Enumeration Areas (EA's).

Since this study fits the framework developed in the previous section, the sample data were taken to illustrate the use of these estimators in a particular application. It should be noted however, that the figures presented are subject to fairly high sampling variability since they are based on a very small cluster sample. The specific estimators used will now be described.

Suppose that the total population is divided into K enumeration assignments. Let M_k be the size of the kth EA and let $\bar{M}$ be the average size of the EA's. Now suppose a simple random sample of k_o EA's is selected from the total, K, and that within each EA all units are observed. Let this sample be denoted by s' and suppose that true values are determined for each unit in s'.

Define

$x_{ik}(t)$: observed value of unit i in EA k at trial t

μ_{ik}: true value of unit i in EA k

$e_{ik}(t) = x_{ik}(t) - \mu_{ik}$: response deviation of unit i in EA k at trial t

Then,

$$\bar{x}_{s'}(t)^* = \frac{1}{k_o\bar{M}} \sum_{k\varepsilon s'} \sum_{i} x_{ik}(t), \quad \bar{\mu}_{s'} = \frac{1}{k_o\bar{M}} \sum_{k\varepsilon s'} \sum_{i} \mu_{ik}$$

$$\bar{e}(t) = (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$$

Also,

$$\bar{x}_k(t) = \frac{1}{M_k} \sum_{i=1}^{M_k} x_{ik}(t), \quad \bar{\mu}_k = \frac{1}{M_k} \sum_{i=1}^{M_k} \mu_{ik}$$

An unbiased estimator of the sampling variance $V_{s'}[\bar{x}_{s'}(t) - \bar{\mu}_{s'}]$ for this sample design is given by

$$v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) = v_{s'}(\bar{e}(t))$$

$$= \frac{K-k_o}{K-k_o} \cdot \frac{1}{k_o-1} \sum_{k=1}^{k_o} \left[\sum_{i=1}^{M_k} \frac{e_{ik}(t)}{\bar{M}} - \bar{e}(t)\right]^2$$

$$= \frac{K-k_o}{K(k_o-1)} \cdot \frac{1}{k_o\bar{M}^2} \left\{ \sum_{k=1}^{k_o} \left(\sum_{i=1}^{M_k} e_{ik}(t)\right)^2 - k_o\bar{M}^2\, \bar{e}^2(t)\right\}$$

Substituting this expression into the MSE formula given in equation 2 gives the following unbiased estimator of MSE $[\bar{x}(t)]$.

---

* This estimator was used because:
1) it is unbiased and thus corresponds to $\bar{x}_{s'}(t)$ in the estimation theory
2) the EA's do not vary very much in size

$$\hat{MSE}[\bar{x}(t)] = \frac{1}{K(k_o-1)} \{k_o(K-1)(\bar{x}_{s'}(t)-\bar{\mu}_{s'})^2 - \frac{(K-k_o)}{\bar{M}^2 k_o} \sum_{k=1}^{k_o} M_k^2(\bar{x}_k(t)-\bar{\mu}_k)^2\} \quad (2.3)$$

If, as frequently occurs, $\bar{X}(t)$ represents the Census proportion of the total population in a given category, then both $x_{ik}(t)$ and μ_{ik} are 0-1 variables. The sample for EA k can therefore be split according to the following table of frequencies.

Table 1:

		Census Classification $x_{ik}(t)$		
		1	0	TOTAL
'True' classification μ_{ik}	1	a_k	b_k	a_k+b_k
	0	c_k	d_k	c_k+d_k
	TOTAL	a_k+c_k	b_k+d_k	M_k

A term often used to measure errors in Census classification is the net difference rate, which, for EA k, is defined as follows

$$r_k = \frac{c_k-b_k}{M_k}$$

For the total sample we define the net difference rate r as

$$r = \frac{c-b}{k_o\bar{M}} \quad \text{where } c = \sum_{k=1}^{k_o} c_k, \quad b = \sum_{k=1}^{k_o} b_k$$

The quantity r is identical to $\bar{e}(t)$ and thus provides an unbiased estimate of the bias in the Census statistic, $\bar{X}(t)$.

In terms of r and r_k, equation (3) may be expressed as

$$\hat{MSE}\ [x(t)] = \frac{1}{K(k_o-1)}\ \{k_o(K-1)\ r^2 - \frac{(K-k_o)}{\bar{M}^2 k_o}\ \sum_{k=1}^{k_o} M_k^2\ r_k^2\}$$

If we assume that k_o is large and that $k_o << K$ this expression may be further simplified to

$$\hat{MSE}\ [\bar{x}(t)] = r^2 - \frac{1}{(\bar{M}\ k_o)^2}\ \sum_{k=1}^{k_o} M_k^2\ r_k^2$$

Similarly, the sampling variance $v_{s'}\ (\bar{x}_{s'}(t) - \bar{\mu}_{s'})$ may be expressed in terms of r and r_k and simplified to

$$v_{s'}(\bar{x}_{s'}(t) - \bar{\mu}_{s'}) = \frac{1}{(\bar{M}\ k_o)^2}\ \sum_{k=1}^{k_o} M_k^2\ r_k^2 - \frac{r^2}{k_o}$$

These two expressions can be easily calculated from the above table of frequencies.

The numerical results are summarized in Table 2. These results confirm the original hypothesis that apartments and duplexes were under-estimated in the 1971 Census.

Table 2: Measures of Reliability of Census Statistics on Type of Dwelling

Type of Dwelling	Census[1] Percentage $\bar{x}(t)$x100%	RMSE ($\bar{x}(t)$)	Estimated Bias ($\bar{x}(t)$)	Standard Error of Estimated Bias ($\bar{x}(t)$)	Estimated True Percentage	Standard Error[2] of Estimated True Percentage
	%	%	%	%	%	%
Single detached.	46.4	1.38	1.69	0.98	44.7	0.98
Double	7.6	2.08	2.20	0.70	5.4	0.70
Duplex	8.1	2.06	-2.36	1.17	10.5	1.17
Single attached and row	8.3	0.44	0.62	0.45	7.7	0.45
Apartment	29.2	2.26	-2.42	0.88	31.6	0.88

[1]Total population includes 0.4% mobile dwellings.

[2]Equivalent to the standard error of the bias.

2. APPROACH II

Under this approach we use as a measure of error the deviation of the specific Census figure from the true parameter being estimated. This differs from the usual response error model in as much as no probability model is assumed on observations being made at Census.

Let $\bar{\mu}$ be the population mean being estimated and let $\bar{x}$ be the corresponding Census figure. Then the net error involved in using $\bar{x}$ as an estimate of $\bar{\mu}$ is given by

$$E = \bar{x} - \bar{\mu}$$

Assume that for a sample s' of the population, the true values can be determined. Let $\bar{x}_{s'}$ and $\bar{\mu}_{s'}$ be unbiased estimates of $\bar{x}$ and $\bar{\mu}$ obtained from the sample respectively. Then an unbiased estimator of E is given by

$$\hat{E} = \bar{x}_{s'} - \bar{\mu}_{s'}$$

The above estimate can be used to produce an alternate estimator of $\bar{\mu}$ by correcting $\bar{x}$ as follows:

$$\hat{\bar{\mu}} = \bar{x} - (\bar{x}_{s'} - \bar{\mu}_{s'})$$

for which the sampling variance is given by

$$V(\hat{\bar{\mu}}) = V(\bar{x}_{s'} - \bar{\mu}_{s'})$$

It is easy to show that $\hat{\bar{\mu}}$ will be more reliable than $\bar{\mu}_s$ when $\bar{x}_{s'}$ and $\bar{\mu}_{s'}$ are highly correlated, which is usually expected.

The above can also be extended to sample surveys when true values are known for a sub-sample of the original sample.

4. CONCLUSION

This paper has presented two approaches for measuring the reliability of Census and sample survey data when true values can be determined for a sample or sub-sample of the population. For each approach, a method of correcting the original estimate was also presented.

Although this theory was developed mainly for applications to response error problems in Census, it is also applicable to other types of situations, e.g. coverage errors, coding errors etc. The particular approach would therefore in general depend on the type of error being investigated.

RESUME

Madow (1968) a proposé un schéma d'échantillonnage à deux degrés suivant lequel le biais de réponse peut être éliminé des enquêtes par sondage en obtenant des valeurs "réelles" pour un sous-échantillon de l'échantillon original. Comme c'est souvent le cas aux recensements ou aux enquêtes en cours, les données des sous-échantillons ne servent pas à corriger les estimations de l'enquête principale, mais à évaluer leur fiabilité. Ce document vise d'abord à présenter des méthodes permettant d'obtenir des estimations de fiabilité lorsque les valeurs "réelles" peuvent être établies pour un sous-échantillon d'unités.

REFERENCES

[1] Madow (1965), "On Some Aspects of Response Error Measurement", Proceedings of the ASA, Social Statistics Section, pages 182-192.

THE TELEPHONE EXPERIMENT IN THE CANADIAN LABOUR FORCE SURVEY

R.C. Muirhead, A.R. Gower, F.T. Newton
Household Surveys Development Division

This paper summarizes the results of a telephone experiment conducted in conjunction with the Canadian Labour Force Survey over the period June 1972 to November 1973. Included in the paper is a detailed outline of the purpose and design of the experiment. A discussion of the impact telephone interviewing had on the cost of enumeration, non-response and participation and unemployment rates is given. In addition, interviewer and respondent attitudes toward telephone interviewing are described. Finally, the paper summarizes the experiences gained from this experiment and indicates some areas where further examinations related to telephone interviewing can be carried out.

1. INTRODUCTION

The Canadian Labour Force Survey is conducted as a monthly sample survey. In one particular week each month about 30,000 households throughout the country are visited by approximately 750 interviewers who obtain information on the labour force activities of all members of each household fourteen years of age and over. All selected households remain in the survey for a period of six months.

In the past the Labour Force Survey used only the face-to-face method of interviewing even though the use of the telephone appeared to have definite advantages, particularly from a cost and timeliness point of view. However, possible repercussions on the quality of the survey were of sufficient consequence to rule out implementation of telephone interviewing until adequate testing and control were carried out. Consequently, beginning in 1971 and continuing through 1972 and 1973, telephone interviewing was introduced on an experimental basis in selected metropolitan areas. Of interest, in addition to the results of this experiment is the gradual and closely monitored expansion of the telephone procedure in order to protect the continuity and validity of the ongoing survey. As a result of the experimental work, telephone interviewing has now been successfully introduced in almost all Canadian cities.

This paper deals with the development and results of the telephone experiment over a period of eighteen months from June 1972 to November 1973. A comprehensive set of tables and graphs highlighting the major findings of the experiment is included. Section 2 details the purpose for which the telephone experiment was established, while Section 3 outlines the design of the experiment. Sections 4, 5 and 6 look at the effect of telephone interviewing on the cost of enumeration, non-response, and unemployment and participation rates. Section 7 discusses interviewer and respondent attitudes toward telephone interviewing. Finally, Section 8 ponders some future considerations.

2. PURPOSE OF THE TELEPHONE EXPERIMENT

During the last few years escalating costs have led to annual increases in the amount of funding required to carry out the Labour Force Survey. It was primarily in an endeavour to offset and to more effectively control these rising costs that an experimental procedure involving the use of telephone interviewing was first suggested. The basic assumptions made were that telephone interviewing would lead to a considerable saving in the cost of enumeration and to some reduction in non-response. The telephone experiment was set up to test these assumptions, and it was designed to provide a measure of the changes in both cost and non-response as well as to measure the effect of telephone interviewing on labour force characteristics such as unemployment and participation rates.

Initially a pilot study was undertaken in 1971 in the Toronto and Vancouver metropolitan areas to determine the feasibility of telephone interviewing and to test the adaptability of interviewers to a telephone operation. This study was operative for four months only, but subsequent analysis of the results indicated that further experimentation should be undertaken. Accordingly, in June 1972, the telephone experiment began in the six English-speaking regional office cities[1] and was further expanded to the two French-speaking regional office cities[2] in April 1973.

---

1. St. John's, Halifax, Toronto, Winnipeg, Edmonton, Vancouver.
2. Montreal, Ottawa-Hull.

3. DESIGN OF THE TELEPHONE EXPERIMENT

3.1 The Telephone and Control Subsamples

In order to compare the telephone interviewing procedure with the regular face-to-face interviewing procedure, a "telephone" subsample and a "control" subsample were selected within the metropolitan area of each regional office city on the basis of interviewer assignments consisting of 45 to 55 households. Interviewers in the telephone subsample were to complete all telephone calls using their own home telephones. Consequently, prior to choosing the subsamples, interviewers who had only party line telephones were excluded from the selection scheme in order to comply with the secrecy provisions of the Statistics Act, and their assignments formed the "non-participant" group. The remaining assignments in each metropolitan area were numbered in a serpentine fashion and systematically allocated from a random starting point to either the telephone subsample or the control subsample. In other words, a circular systematic subsampling of interviewer assignments was made, so that about one half the assignments were selected for the telephone subsample and one half for the control subsample.

Thus, the design of the telephone experiment yielded two major categories: (a) the telephone subsample and (b) the control subsample. Although this design did not remove different interviewer effects between the subsamples, it was easy to implement and control in the field. Moreover, the effect of rotation group bias was minimal since all rotations groups had an equal representation in both subsamples.

3.2 Interviewing Procedure

In the telephone subsample, interviewers were required to make both personal visits and telephone calls. Any first month interview[1] with a household was to be completed by a personal visit. If the respondent granted permission to be interviewed by telephone, then the subsequent

1. First month interviews include situations when (a) a dwelling containing a household rotates into the Labour Force sample, (b) a dwelling which was vacant in the previous month is now occupied by a household, and (c) there has been a complete change in the composition of the household.

months' interviews were conducted using the telephone. Personal visits also had to be made to households which could not be telephoned for any of the following reasons: no telephone available, refusal to be interviewed by telephone, telephone interview impossible due to language or hearing problems, and inability to reach the respondent by telephone (personal follow-ups were required to complete an interview or to determine the reason for no interview).

Interviewers in the control subsample, on the other hand, made only personal visits to households in their assignments.

3.3 Implementation of the Telephone Experiment

In June 1972 telephone and control subsamples were established in the metropolitan areas of six regional office cities: St. John's, Halifax, Toronto, Winnipeg, Edmonton and Vancouver. Telephone and control subsamples were also set up in the regional office cities of Montreal and Ottawa-Hull in April 1973. The division of the metropolitan areas of St. John's, Montreal, Toronto and Vancouver into telephone and control subsamples continued until November 1973. At that time results indicated the desirability of the telephone procedure. Therefore, in December 1973 all interviewers in the control subsample began telephone interviewing. Earlier, in March 1973, all control assignments in Edmonton had been designated for telephone interviewing, but the two subsamples continued in Halifax, Ottawa-Hull and Winnipeg until March 1974. Gradual introduction of telephone interviewing to cities other than regional office centres has been in progress since December 1973.

In this paper monthly results are frequently presented for all regional office cities combined. Because the experimental phase-in of the eight cities occurred at different times, the data at this level represents various combinations of cities depending on the month as follows:

(a) June 1972 to February 1973 - St. John's, Halifax, Toronto, Winnipeg, Edmonton, Vancouver

(b) March 1973 - St. John's, Halifax, Toronto, Winnipeg, Vancouver

(c) April 1973 to November 1973 - St. John's, Halifax, Montreal, Ottawa-Hull, Toronto, Winnipeg, Vancouver.

4. THE COST OF ENUMERATION

The cost of enumerating households consists of two components: (a) fees which represent the total cost of the time taken for interviewing households, contacting households (travelling time, making call-backs, and so on) as well as completing work for transmittal and home study exercises, and (b) expenses which include the mileage cost of travelling and other authorized expenses such as meals. The sum of fees and expenses gives the enumeration cost. The enumeration cost per household is calculated by dividing the enumeration cost by the number of households (all sampled dwellings less the number of vacant dwellings). Fees per household and expenses per household are calculated in a similar manner.

In order to measure the savings achieved by the telephone procedure, the percentage difference between the cost (fees, expenses, enumeration cost) per household in the telephone subsample and in the control subsample was calculated each month. This measure, called the percentage saving in cost per household for the telephone subsample over the control subsample, is defined as $(C_c - C_t)/C_c \times 100\%$ where C_c and C_t denote the cost per household in the control and telephone subsamples respectively.

The percentage savings in fees, expenses and enumeration cost per household for individual regional office cities are summarized on Table 4.1. This table shows that the telephone interviewing procedure led to a substantial saving in the cost of enumeration over the personal visit interviewing procedure. Although it was impossible to identify the enumeration cost associated with telephone calls only, it is reasonable to assume that the enumeration cost of telephone interviewing was lower than face-to-face interviewing since the amount of time spent on travelling was reduced (indicated by the substantial saving in the expenses component).

While the cost of first month visits in both subsamples can be assumed to have been approximately equal, the cost of completing other personal visits within the telephone subsample was probably very high due to the long distances necessary to complete only a few households. The savings which occurred in the telephone subsample, therefore, appear to have resulted from the large number of telephone interviews. Table 4.2 summarizes the enumeration cost per household in the telephone and control subsamples for all regional cities combined over the one year period from December 1972 to November 1973.

Table 4.1: Percent savings in enumeration cost per household between telephone and control subsamples by regional office city, December 1972 to November 1973

R.O. City	Fees	Expenses	Enumeration Cost
St. John's	2.2%	32.0%	8.5%
Halifax	17.6%	48.1%	22.8%
Montreal[1]	20.7%	34.9%	23.1%
Ottawa-Hull[1]	5.0%	4.3%	4.8%
Toronto	17.5%	37.2%	20.9%
Winnipeg	- 9.9%	28.9%	- 0.5%
Edmonton[2]	26.0%	50.0%	29.8%
Vancouver	11.5%	20.0%	12.8%
All R.O. cities	13.7%	32.5%	17.4%

1. April 1973 to November 1973 only.
2. December 1972 to February 1973 only.

Table 4.2: Enumeration cost per household in the telephone and control subsamples, all regional office cities combined, December 1972 to November 1973

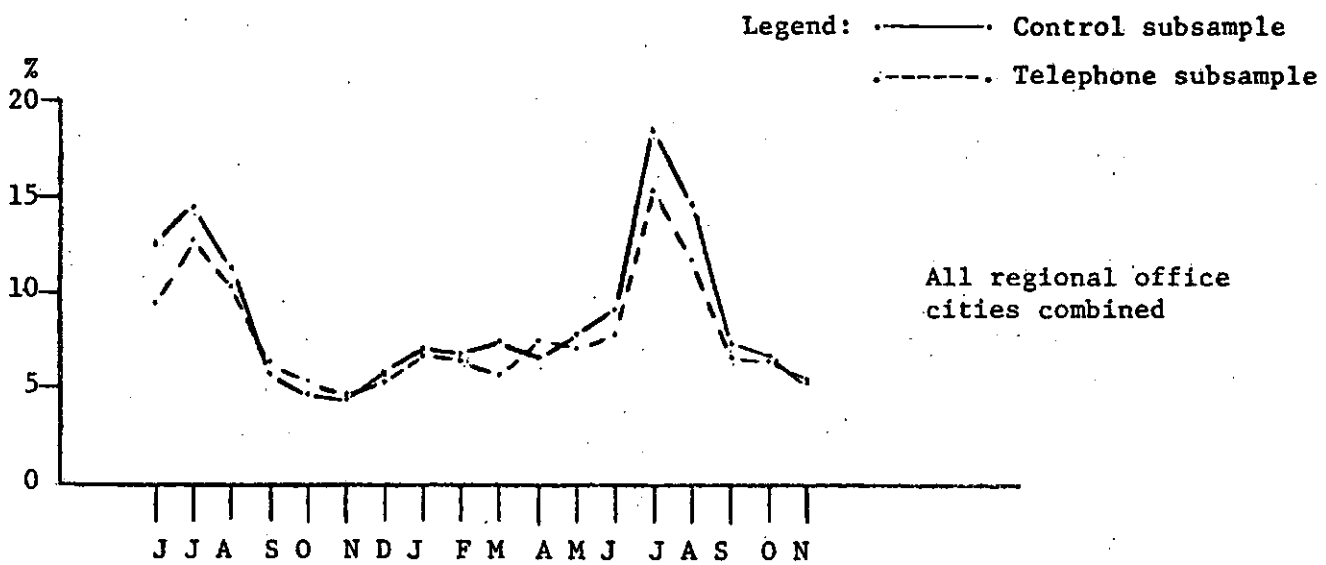
Month	Enumeration cost per household		Percentage saving telephone over control
	Telephone subsample	Control subsample	
December	$1.74	$2.12	17.9%
January	$1.83	$2.15	14.9%
February	$1.68	$2.04	17.6%
March	$1.68	$2.15	21.9%
April	$1.61	$1.93	16.6%
May	$1.86	$2.23	16.6%
June	$1.82	$2.17	16.1%
July	$1.87	$2.17	13.8%
August	$1.85	$2.27	18.5%
September	$2.01	$2.47	18.6%
October	$2.11	$2.56	17.6%
November	$1.98	$2.49	20.5%

5. NON-RESPONSE

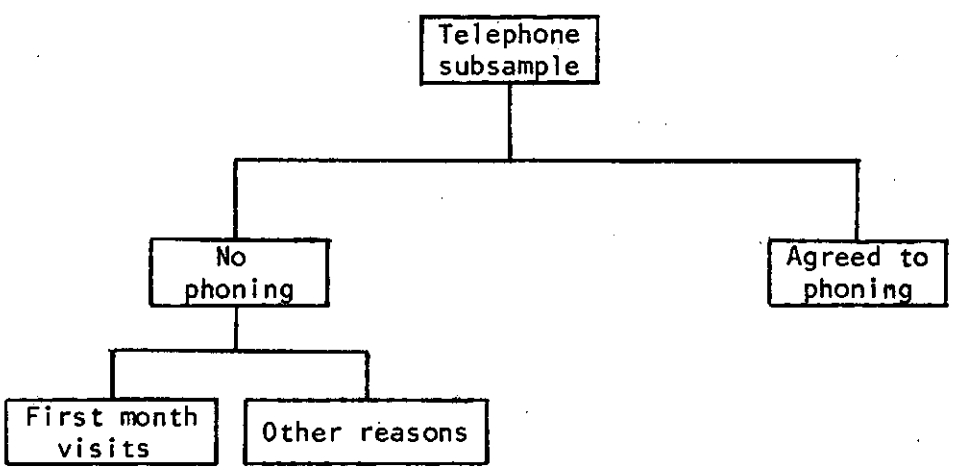
In any survey non-response can be expected to occur due to operational difficulties, time and cost restraints, the lack of co-operation from respondents, the inability or unwillingness of interviewers to work hard enough to track down missing respondents, or for some other reason. The non-response rate measures the severity of this non-response problem, and it is calculated as the percentage of non-respondent households out of all sampled households.

Non-response rates were calculated each month for both the telephone and control subsamples, and these rates are shown on Graph 5.1 for all cities combined and Graph 5.2 for each regional office city.

Graph 5.1: Non-response rates in the telephone and control subsamples, June 1972 to November 1973

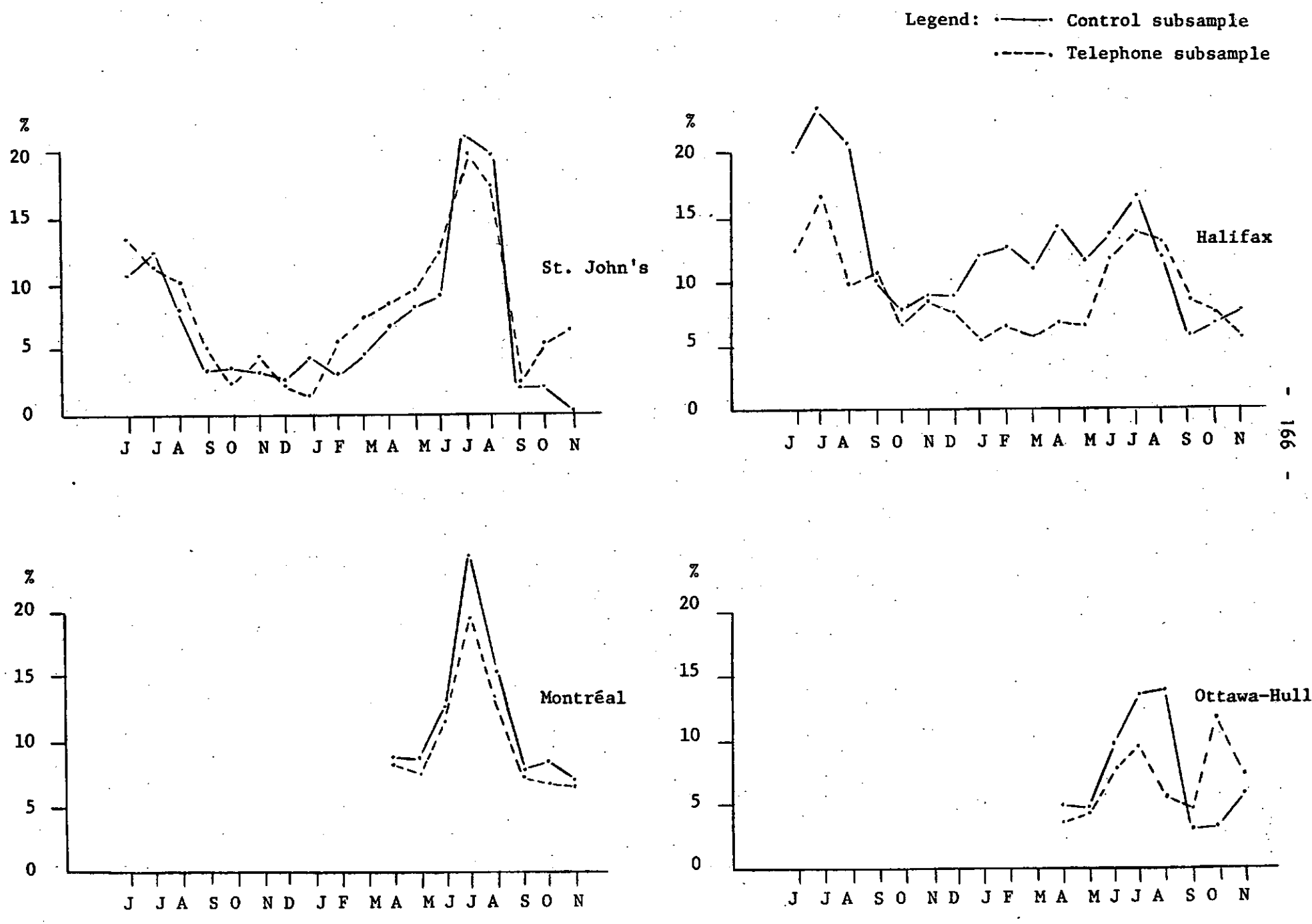


Although it had originally been hypothesized that the telephone interviewing procedure would lead to some reduction in non-response, this reduction does not appear to have occurred. However, the telephone experiment made possible a more detailed analysis of non-response by partitioning the telephone subsample into the following groups.

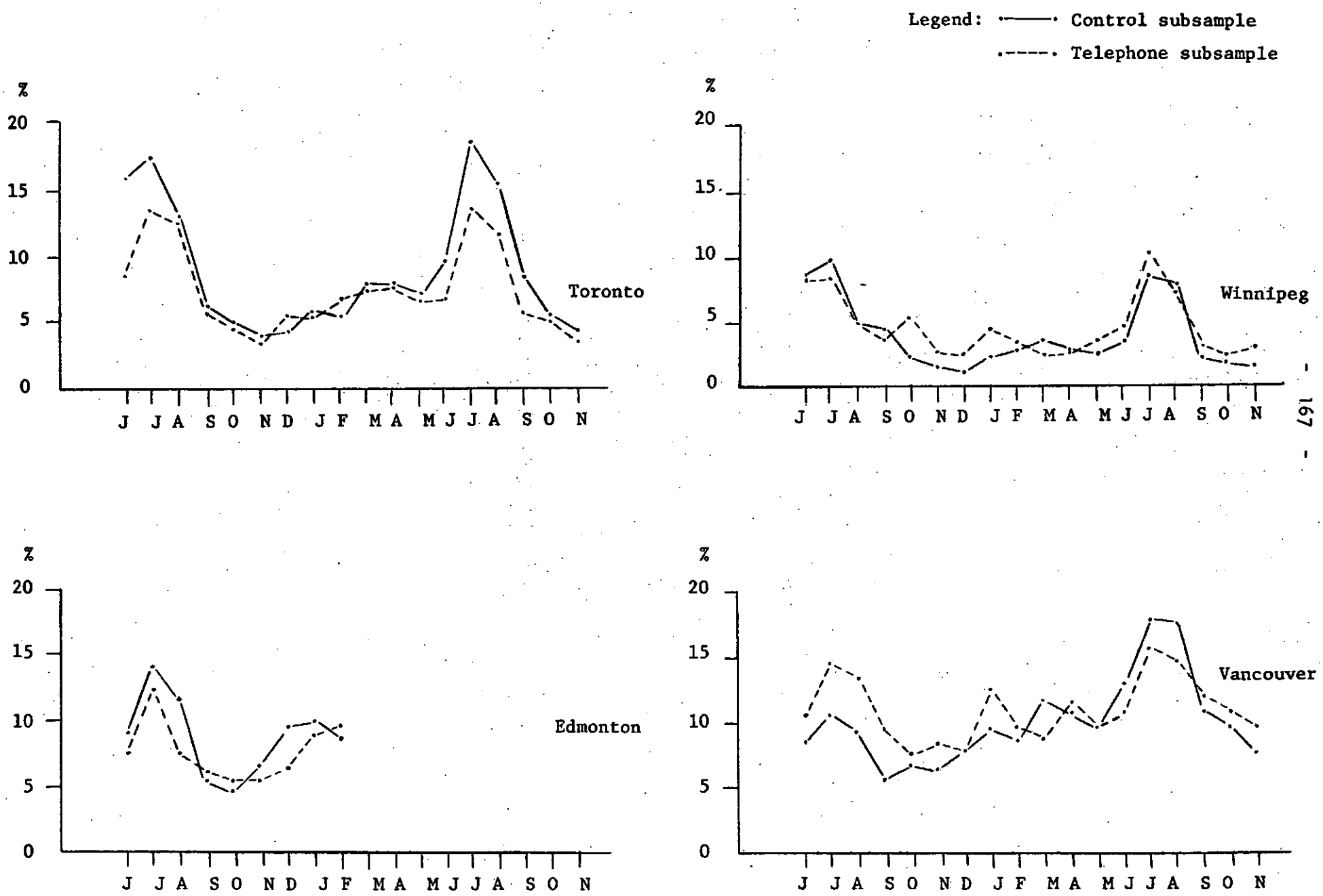


In this diagram, the "agreed to phoning" group is the aggregate of all households which agreed to be interviewed by telephone and received at least one phone call during interview week. The "no phoning" group

Graph 5.2: Non-response rates in the telephone and control subsamples, June 1972 to November 1973



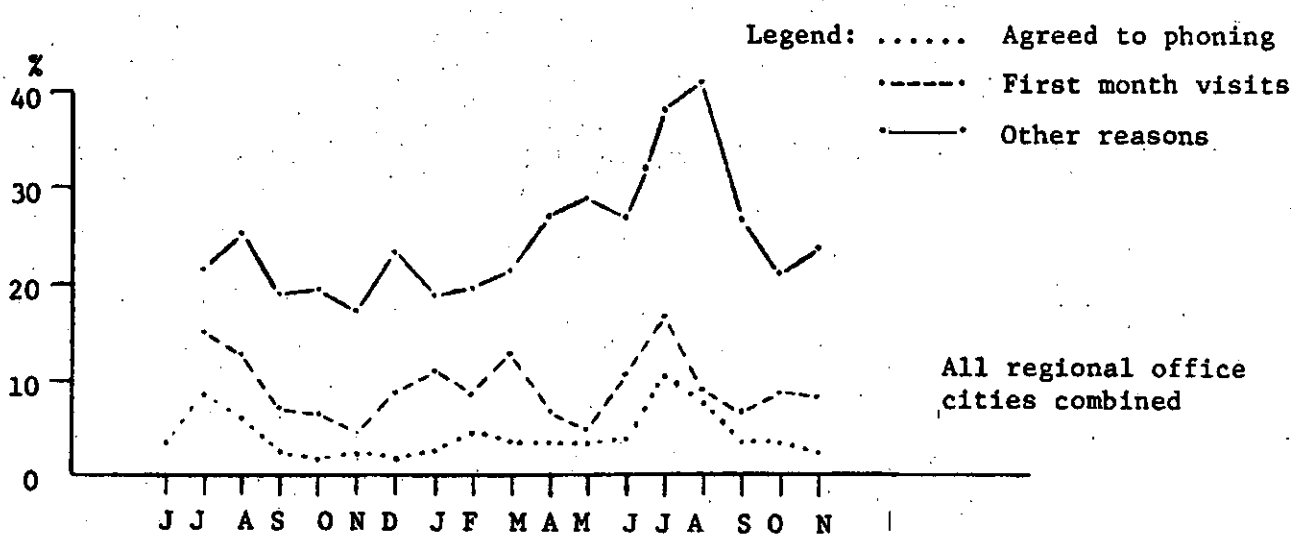
Graph 5.2: Non-response rates in the telephone and control subsamples, June 1972 to November 1973



consists of two sub-groups: "first month visits" and "other reasons". The "first month visits" group includes households which rotated into the Labour Force sample for the first time as well as households which, although in the sample, had not been enumerated in previous months. The "other reasons" group consists of all remaining households in the telephone subsample which were not telephoned for reasons such as refusal to give consent, no telephone available, or respondent had hearing or language problems.

The telephone subsample, therefore, was able to be partitioned into three groups: (a) agreed to phoning, (b) first month visits, and (c) other reasons. Graph 5.3 illustrates for all regional office cities combined the relationship that existed among the non-response rates for these three groups.

Graph 5.3: Non-response rates in the agreed to phoning, first month visits, and other reasons groups in the telephone subsample, June 1972 to November 1973



It can be seen from this graph that the households in the "agreed to phoning" group were co-operative, and the results reflect this situation since the non-response rate for this group was lower than the other two groups. The non-response rate in the "first month visits" group was higher than the "agreed to phoning" group but substantially lower than the "other reasons" group.

Table 5.1 shows the percentage of households belonging to each group as well as the contribution to the total non-response by each group. The households in the "other reasons" group, on the average, had a non-response rate of approximately 25 percent and accounted for almost 45 percent of all non-respondents in the telephone subsample even though they represented only 15 percent of all households in the subsample. Thus, a small group of households accounted for a very large portion of non-response. In this way, the telephone interviewing procedure has provided a means of isolating the group of households which will have the greatest probability of being non-respondents. These households can be identified after the first month interview. Since they cannot be interviewed by telephone for one reason or another, the interviewer will be forewarned that they will require an extra effort to obtain successful interviews. In the long run this should result in lower non-response rates as interviewers acquire special skills in handling this type of problem household.

Table 5.1: Average contribution to non-response by the agreed to phoning, first month visits, and other reasons groups in the telephone subsample, July 1972 to November 1973

Regional office city	Number of households	Number of non-respondent households	Group	Percentage of households	Contribution to non-response
St. John's	135	11	Agreed to phoning	74.4%	41.2%
			First month visits	17.3%	18.7%
			Other reasons	8.3%	40.1%
Halifax	260	24	Agreed to phoning	71.0%	33.9%
			First month visits	16.5%	18.7%
			Other reasons	12.5%	47.4%
Montreal[1]	1,223	124	Agreed to phoning	68.5%	38.3%
			First month visits	18.0%	19.3%
			Other reasons	13.5%	42.4%
Ottawa-Hull[1]	271	18	Agreed to phoning	58.0%	34.9%
			First month visits	17.7%	21.9%
			Other reasons	24.3%	43.2%
Toronto	1,180	87	Agreed to phoning	68.7%	40.9%
			First month visits	17.3%	19.6%
			Other reasons	14.0%	39.5%
Winnipeg	505	22	Agreed to phoning	66.5%	39.8%
			First month visits	17.2%	23.9%
			Other reasons	16.3%	36.3%
Edmonton[2]	423	32	Agreed to phoning	70.4%	26.9%
			First month visits	15.3%	21.3%
			Other reasons	14.3%	51.8%
Vancouver	615	64	Agreed to phoning	65.7%	26.2%
			First month visits	17.3%	19.8%
			Other reasons	17.0%	54.0%
All regional office cities combined			Agreed to phoning	67.9%	35.5%
			First month visits	17.3%	20.0%
			Other reasons	14.8%	44.5%

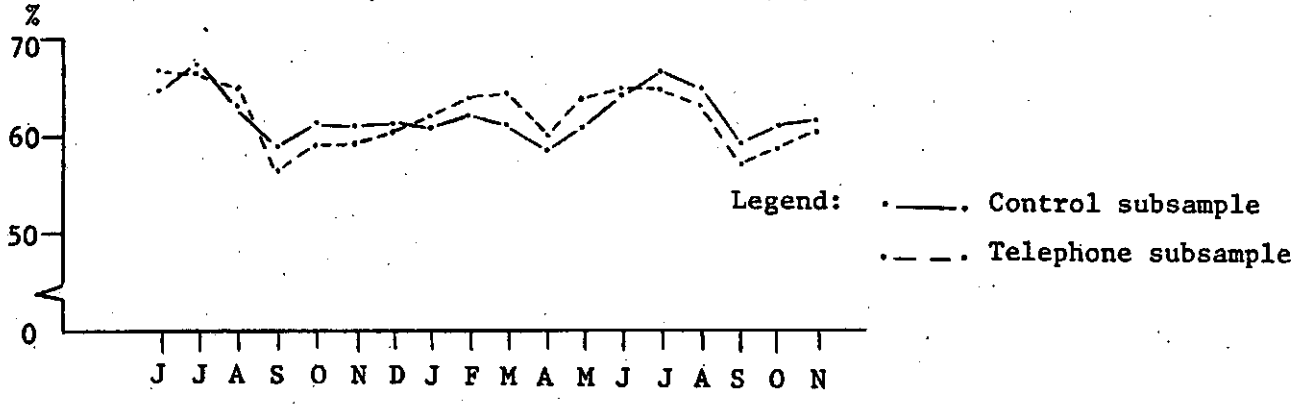
1. April 1973 to November 1973 only.
2. June 1972 to February 1973 only.

6. PARTICIPATION AND UNEMPLOYMENT RATES

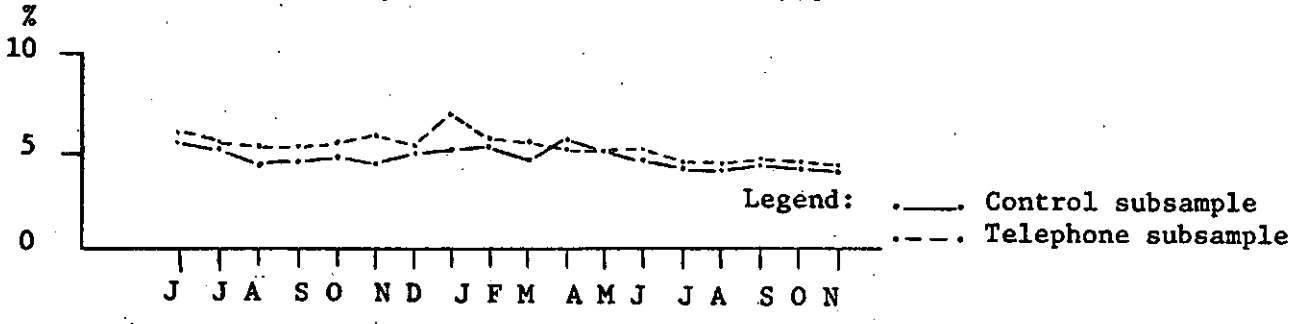
An important aspect of the telephone experiment was to study the effect of the telephone interviewing procedure on labour force characteristics. Participation and unemployment rates were examined in detail. The participation rate represents the labour force[1] as a percentage of the civilian non-institutional population fourteen years of age and over, while the unemployment rate is the number of unemployed persons as a percentage of the labour force.

Graphs 6.1 and 6.2 below show the participation and unemployment rates in the telephone and control subsamples for all regional office cities combined from June 1972 to November 1973.

Graph 6.1: Participation rates, all regional office cities combined, June 1972 to November 1973



Graph 6.2: Unemployment rates, all regional office cities combined, June 1972 to November 1973



1. The labour force is the civilian non-institutional population fourteen years of age and over who, during the reference week for the Labour Force Survey, were employed or unemployed.

Statistical tests of significance were carried out for each month, for all regional office cities combined, and it was found that no significant differences occurred between the participation rates in the telephone and control subsamples. With respect to unemployment rates, the results indicate that the unemployment rate in the telephone subsample remained at a higher level than the unemployment rate in the control subsample for sixteen months of the eighteen month experimental period, but the observed differences were statistically significant in only two months: November 1972 and January 1973. However, it was notable that the unemployment rates in each subsample generally followed the same pattern over the eighteen months and moved in the same direction from month to month.

In order to learn more about the differences in the unemployment rates between the telephone and control subsamples, tests of significance were also undertaken for each regional office city. Significant differences did not occur more than twice in every city except Winnipeg. In Winnipeg, the unemployment rate in the telephone subsample was greater than the corresponding rate in the control subsample in seventeen out of eighteen months and significant differences between the two subsamples occurred in six of these months. A special study was carried out to explore the possibility of an "area effect" on labour force characteristics in that the telephone subsample in the Winnipeg metropolitan area was biased toward a less affluent (and hence higher unemployment) area. The results of the study substantiated this possibility, and it was concluded that the unemployment rates in the two subsamples in Winnipeg had different expected values and that the selection scheme had biased the results rather than the interviewing procedure itself.

After April 1973, the unemployment rates in the telephone and control subsamples tended to converge and became almost equal. This appears to have occurred because Montreal was introduced to the experiment at this time, and in Montreal as well as Toronto (where the labour force population was very large in each subsample) the unemployment rates in the two subsamples followed the same general pattern and were almost equal throughout the experiment.

The conclusion was reached that the telephone methods of interviewing did not adversely affect the measurement of labour force characteristics. Moreover, through the gradual and closely monitored introduction of the telephone procedure to other cities, it was felt that the continuity and validity of the ongoing survey was protected.

7. INTERVIEWER AND RESPONDENT ATTITUDES TOWARD TELEPHONE INTERVIEWING

7.1 Interviewer Attitudes

The interviewer's job is to collect accurate information on labour force status and occasionally supplementary subjects from a sample of households and persons. The final statistics can be no better than the information recorded at this initial stage and, therefore, a great deal of faith and reliance has to be placed on the interviewer's ability to gain the confidence of the respondent. Does the telephone facilitate this work? Does the fact that personal contact is reduced affect an interviewer's attitude to the new procedure? In order to answer these questions interviewers were asked in October 1972 to complete a questionnaire designed to pick up their reactions to the telephone procedure, and on two later occasions group sessions were held to discuss their concern and their feelings about the new interviewing procedure.

It was found that many interviewers take the job because it gives them a chance to get out of the house, to meet other people, and at the same time to do something useful while earning money. The telephone procedure initially created some hostility in interviewers as they felt it would cut their chances of getting out and meeting other people. These feelings were responsible for a less than satisfactory performance during the early months but as the realization took hold that almost 30 percent of households still required a personal visit (first month visits, households which denied permission to be interviewed by telephone, call-backs, etc.) much of this hostility vanished, and during the winter months telephoning became a positive boon since interviewing could be completed on time regardless of weather conditions.

As interviewers gained facility in the new procedure they indicated that, given the present mix of telephone and personal visits, they could handle assignments of 70 to 90 households compared to 45 to 55 households prior to the introduction of telephone interviewing. However, not all interviewers were happy with the change of method. A check carried out early in the experiment on assignments of interviewers who either did not like or were not happy with the telephone procedure indicated these assignments contained a large proportion of highly mobile persons (for example, immigrants and other persons living mostly in collective-type dwellings such as rooming and boarding houses). These situations did not lend themselves to the use of the telephone, as frequently only one telephone was available and it was located in a common hall. Therefore, personal visits were necessary in these cases.

A general impression from the frequent questioning of interviewers supported the theory that the longer a procedure is used the more it gains in acceptability. As interviewers became more adept at using the new procedure, the less fault they found with it. Interviewers hired after the introduction of the technique have shown no reservations whatsoever.

7.2 Respondent Attitudes

Respondent acceptance was measured by calculating the rate of telephone activity. This measure was defined as the percentage of households interviewed on the telephone out of all respondent households in the telephone subsample. Ideally this rate should include all households except the one-sixth that rotate into the Labour Force sample each month (that is, in percentage terms about 83 percent of all interviewed households).

The extent of this acceptance was shown quite dramatically in the difference between the overall results for June 1972 (introductory month) and November 1973 (termination of review period). The actual rate of telephone activity increased from 59 percent to 74 percent and, given that first month visits were relatively stable in both months at about 17 percent,

the reduction in the percentage of households visited personally in the telephone subsample decreased from about 24 percent in June 1972 to less than 9 percent in November 1973. This represents a very remarkable improvement over a span of seventeen months.

Table 7.1 provides information on the rate of telephone activity. Column 3 on this table indicates the percentage of interviewed households actually completed by telephone, and column 4 provides similar data for households which should have been telephoned but which, for one reason or another, were enumerated by personal visit. Columns 5 to 10 give the breakdown by reason of the personal visits reported in column 4. Table 7.2 provides similar information for a post survey period of four months. It may be significant to the possible future expansion of telephone usage that, while the proportion of telephone interviews to personal visits in Table 4.1 was roughly two-thirds telephone and one-third personal visits, corresponding percentages in Table 7.2 were almost three-quarters telephone and one-quarter personal visits. In other words, this seems to suggest that the rate of telephone activity can be expected to stabilize at 74 or 75 percent, with 17 percent first month visits and only 7 or 8 percent requiring personal visits for any other reason.

Table 7.1: Average rates of telephone activity in the telephone subsample, June 1972 to November 1973

Regional office city	Interviewed households			Personal visits by reason (%)					
	Average households per survey	By telephone (%)	By personal visit (%)	First month visit	Permission to phone denied	No telephone available	Language or hearing problems	Complete change in household composition	Other
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
St John's	124	75.1	24.9	17.0	0.3	1.9	0.0	1.1	4.6
Halifax	236	72.2	27.8	16.8	1.7	3.4	0.2	2.4	3.3
Montreal[1]	1,099	70.5	29.5	18.1	1.9	2.3	1.4	1.5	4.3
Ottawa-Hull[1]	253	55.7	44.3	17.5	2.1	2.9	0.6	1.3	19.9
Toronto	1,093	67.9	32.1	17.1	2.5	1.9	2.9	2.1	5.6
Winnipeg	483	64.9	35.1	17.0	10.8	2.6	1.0	1.7	2.0
Edmonton[2]	391	71.2	28.8	15.6	2.5	2.3	1.0	5.0	2.4
Vancouver	551	69.7	30.3	17.4	1.5	4.0	1.9	2.3	3.2
All R.O. cities		68.5	31.5	17.2	3.3	2.6	1.7	2.1	4.6

1. April 1973 to November 1973 only.
2. June 1972 to February 1973 only.

Table 7.2: Average rates of telephone activity, December 1973 to March 1974

Regional office city	Interviewed households			Personal visits by reason (%)					
	Average households per survey	By telephone (%)	By personal visit (%)	First month visit	Permission to phone denied	No telephone available	Language or hearing problems	Complete change in household composition	Other
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
St. John's	325	77.8	22.2	17.1	0.2	2.8	0.0	1.0	1.1
Halifax	255	79.8	20.2	15.2	0.4	2.0	0.0	1.4	1.2
Montreal	2,278	74.6	25.4	17.8	1.4	2.4	0.4	0.4	3.0
Ottawa-Hull	412	59.4	40.6	17.0	1.1	1.9	0.9	1.0	18.7
Toronto	927	73.8	26.2	18.2	1.6	1.6	1.7	0.5	2.6
Winnipeg	499	77.5	22.5	16.5	1.5	2.0	1.1	1.0	0.4
Edmonton[1]	-	-	-	-	-	-	-	-	-
Vancouver	1,169	72.6	27.4	18.6	1.0	3.2	2.0	0.8	1.8
All R.O. cities		73.8	26.2	17.7	1.2	2.4	1.0	0.7	3.2

1. No data available for Edmonton.

8. FUTURE CONSIDERATIONS

The telephone experiment was, for those of us involved, a unique experience because no previous data was known to be available in Canada or elsewhere to guide our deliberations or to provide a degree of support for the decisions we took and the actions we embarked upon. We learned, therefore, as we progressed, and in the process we acquired two things: firstly, a considerable expertise in a telephone interviewing procedure and secondly, a knowledge of the direction in which our future studies should lead us.

As a consequence of the information gained during the experimental period, a phased introduction of telephoning in other large cities was initiated. This was carried out in a single step; that is, an entire city was converted to telephone interviewing without recourse to the intermediate telephone and control observation period. For several months after each introduction, a monitoring procedure was set up which enabled an immediate study of key indicators such as non-response and cost to be carried out. This system allowed quick action whenever problem areas surfaced.

These new cities behaved in a similar manner to the original eight regional office cities. For example, nine cities introduced to telephoning during December 1973 and January 1974 had, over a three month period, achieved rates of telephone activity of approximately 74 percent. This was almost identical with the rate for the original eight cities over the same period. Non-response and cost also followed a similar parallel to the eight original cities.

It would seem, therefore, that the "settling in" period is over and telephone behavior can now be predicted within close limits. The difficulty of interviewing, for example, may be determined from a respondent's willingness to accept or reject a telephone interview. If a household is agreeable, the risks of refusal in the future are less and the cases of non-response due to absence are minimized by a more flexible timetable of phone-backs. Households which cannot be telephoned, however, require special skills and attention. In fact, our experience has shown that many of these households have never completed an interview. A unique feature of the

telephone experiment was that it provided a means of isolating these households for further study. These results provoked considerable thought, and one of the outcomes was the belief that perhaps a centralized telephone interviewing procedure would enable these so-called problem households to be dealt with more effectively.

The basis for this centralized strategy would be a pairing concept whereby each assignment would be assigned two interviewers: one to specialize in personal visiting and the other to specialize in telephone interviewing only. The former would be located in the area of the assignment while the latter could be located at a position remote from the assignment. There are many additional consequences that could be dwelt upon. However, what is clear is that centralized telephoning should streamline the overall interview operation and allow for greater specialization in techniques. It is to this area that our further examinations will lead us.

One rather important consequence of the increased use of the telephone is that telephoning makes it more difficult to conduct lengthy supplementary surveys. This serious problem has not escaped attention, to the extent that work on the development of an overall supplementary strategy is underway.

Has the telephone project been successful? Was the experimentation worth the time, effort and money spent upon it? If usage of the procedure is the criteria, then it has been very successful. Coverage of the telephone procedure in the ongoing Labour Force Survey has expanded to include approximately 42 percent of the households in all self-representing and non-self-representing units combined, or about 76 percent of the households in self-representing units. If cost reduction is a criteria, then the experiment must again be termed successful. Moreover, in the long run, the ability to predict and to isolate problem households for non-response will enable corrective measures to be introduced. This should lead to higher response rates.

9. ACKNOWLEDGMENT

The authors wish to thank the referee for some helpful comments.

RESUME

Le présent document résume les résultats d'une expérience téléphonique menée conjointement avec l'enquête sur la population active canadienne pour la période de juin 1972 à novembre 1973. Les buts et le plan de l'expérience sont exposés en détail. On discute de l'incidence des entrevues par téléphone sur le coût du dénombrement, les taux de non-réponses et de participation et sur les taux de chômage. De plus, on décrit l'attitude des interviewers et des répondants envers ces entrevues. Enfin, l'article résume les conclusions tirées de l'expérience et indique certains domaines qui pourront faire l'objet d'études supplémentaires relatives aux entrevues par téléphone.

REFERENCES

[1] Gower, A.R., "Enumeration Cost of Personal Visit and Telephone Assignments in the Labour Force Survey", LFS-DP-12, October 1974.

[2] Gower, A.R., "Preliminary Report on Centralized Telephone Interviewing", LFS-74-110, January 1974.

[3] Newton, F.T., "The Labour Force Survey Telephone Experiment, June to October 1972", LFSP-73-65, December 1972.

[4] Tourigny, J.Y., "Telephone Experiment: Non-Response Rates", LFSP-73-101, November 1973.

[5] Tourigny, J.Y., "Telephone Experiment: Rate of Telephone Activity", LFSP-73-86, July 1973.

ON THE IMPROVEMENT OF SAMPLE SURVEY ESTIMATES

V. Tremblay
Household Surveys Development Division

This paper focuses on the improvement of sample survey estimates in the particular situation where the survey sample, or part of it, is included in a larger sample from which auxiliary information is available. The properties of a method of estimation - sometimes applied in specific circumstances - are investigated and the limitations of its application are found. The application of the method to rotation designs in continuing surveys is more closely studied in the context of composite estimation.

1. INTRODUCTION

The present paper studies an estimation procedure which takes into consideration the information collected in the first phase for the purpose of improving the estimates derived in a second phase. This estimation procedure was first examined by Tenenbein [1] in the context of correction of misclassified multinomial data. A more general treatment is provided in this paper. In the first section, the procedure is described, and some properties of the estimates are given. The efficiency of the procedure is demonstrated in the second section. Finally, the application of the method for building up composite estimates for continuing surveys is examined in the last section.

The procedure as presented may be applied to several other situations such as 1) to make survey data consistent with some census data 2) to improve estimates derived from multiphase surveys 3) to make use auxiliary or administrative data to complement survey data and 4) to correct misclassified multinomial data, as Tenenbein proposed it.

2. ESTIMATION PROCEDURE

2.1 The General Problem

A population P is being surveyed twice. In phase 1, a sample S is selected to estimate the partition $X = (X_1, X_2, \ldots, X_p)'$ of P where X_j is the number of units of P having the characteristic C_j with $C = \{C_j: j=1, \ldots, p\}$ being a set of mutually exclusive characteristics such that each unit of

P has one and only one of the characteristics of C. In phase 2, a subsample $S_1 \subset S$ is selected to estimate the partition $Z = (Z_1, Z_2, \ldots, Z_q)'$ of P where Z_i is the number of units of P having the characteristic C_i' with $C' = \{C_i' : i = 1, \ldots, q\}$ being a set of mutually exclusive characteristics such that each unit of P has one and only one of the characteristics of C'. Furthermore, for each unit of the sample $S_1 \subset S$ it is possible to know what its C characteristic is.

We shall assume in this paper that both samples S_1 and S_2 have been obtained through simple random sampling drawn without replacement and that S is the union of these two independent samples. An attempt was made, however, to develop proofs in their most general forms, so that these could be adapted to other sampling designs. Some references are made to situations where S is a complete census, since such cases are currently applied; however then, S can no longer be assumed to be the union of two independent samples.

Splitting the sample S into two independent components S_1 and $S_2 = S \sim S_1$, it will be convenient to describe the estimation procedure according to the following diagram:

	Sample S_1	Sample $S_2 = S \sim S_1$
Phase 1	$\hat{X}^{(1)} = \begin{bmatrix} \hat{X}_1^{(1)} \\ \vdots \\ \hat{X}_p^{(1)} \end{bmatrix}$	$\hat{X}^{(2)} = \begin{bmatrix} \hat{X}_1^{(2)} \\ \vdots \\ \hat{X}_p^{(2)} \end{bmatrix}$
Phase 2	$\hat{Z}^{(1)} = \begin{bmatrix} \hat{Z}_1^{(1)} \\ \vdots \\ \hat{Z}_q^{(1)} \end{bmatrix}$	

Here, $\hat{X}^{(1)}$ (or $\hat{X}^{(2)}$) and $\hat{Z}^{(1)}$ are the simple weighted estimates of X and Z from samples S_1 (or S_2) with

may be easily proved assuming that the sample S is the union of two independent samples S_1 and S_2 obtained under SRSWR design.

$$E\ \hat{m}_{ij}^{(1)} \mid \hat{X}^{(1)} = m_{ij} . \tag{2.2}$$

$$E\ \hat{m}_{ij}^{(1)} = m_{ij} = R_{ij}/X_j . \tag{2.3}$$

$$\text{Cov}\ \hat{X}_j, \hat{X}_k = \omega\ \text{Cov}\ \hat{X}_j^{(1)}, \hat{X}_k^{(1)} = (1-\omega)\ \text{Cov}\ \hat{X}_j^{(2)}, \hat{X}_k^{(2)} . \tag{2.4}$$

$$\text{Cov}\ \hat{R}_{ij}^{(1)}, \hat{R}_{ik}^{(1)} \mid \hat{X}^{(1)} = 0 \qquad \text{for } j \neq k . \tag{2.5}$$

$$\text{Cov}\ \hat{m}_{ij}^{(1)}, \hat{m}_{ik}^{(1)} = 0 \qquad \text{for } j \neq k . \tag{2.6}$$

$$\text{Var}\ \hat{m}_{ij}^{(1)} = X_j^{-2}\ (1 + \text{Rel Var}\ \hat{X}_j^{(1)})\ E\ \text{Var}\ \hat{R}_{ij}^{(1)} \mid \hat{X}^{(1)} . \tag{2.7}$$

This last equality is the approximation when dropping terms smaller in order of magnitude than the relative variance of $X_j^{(1)}$. Finally,

$$\text{Cov}\ \hat{R}_{ik}^{(1)}, \hat{m}_{ij}^{(1)}\ \hat{X}_j^{(2)} = 0 \qquad \text{if } j \neq k \tag{2.8}$$

$$= E\ \text{Var}\ \hat{R}_{ik}^{(1)} \mid \hat{X}^{(1)} \qquad \text{if } j = k$$

2.3 Properties of the method:

i) $\hat{Z}$ is an unbiased estimator of Z. It is sufficient to prove that $\hat{Z}_1^{(2)}$ is unbiased. Samples S_1 and S_2 being independent, we have

$$E\ \hat{Z}_1^{(2)} = \Sigma_{j=1}^{P}\ E\ \hat{m}_{1j}^{(1)}\ X_j^{(2)} = \Sigma_{j=1}^{P}\ E\ \hat{m}_{1j}^{(1)}\ X_j$$

Now, applying the equation (2.3) along with the definition or R_{1j}, we find

$$E\ \hat{Z}_1^{(2)} = \Sigma_{j=1}^{P}\ R_{1j} = Z_1 .$$

ii) For all i's ($i = 1, \ldots, q$), $\hat{Z}_i$ is maximum likelihood estimate of Z_i. Let's fix $i=1$. Considering the independence of samples S_1 and S_2, the likelihood function of $\hat{R}_{1j}^{(1)}$, $\hat{X}_j^{(1)} - \hat{R}_{1j}^{(1)}$ and $\hat{X}_j^{(2)}$ is proportional to

$$L_1 = \pi_{j=1}^{p} (R_{1j}/T)^{\hat{R}_{1j}^{(1)}} (X_j/T - R_{1j}/T)^{\hat{X}_j^{(1)} - \hat{R}_{1j}^{(1)}} \pi_{j=1}^{p} (X_j/T)^{f.\hat{X}_j^{(2)}}$$

where T is the total population size and $f = (1 - \omega)/\omega$ is the ratio of size of S_2 to the size of S_1. The maximum likelihood estimate of R_{1j} is then found to be

$$= \omega \hat{R}_{1j}^{(1)} + (1 - \omega) (\hat{X}_j^{(2)}/\hat{X}_j^{(1)}) \hat{R}_{1j}^{(1)} ,$$

and the results follows from the fact that

$$\Sigma_{j=1}^{p} R_{1j} = Z_1 .$$

iii) The estimators obtained using this method are consistent with those derived from sample S in the first phase; explicitly we have:

$$\Sigma_{i=1}^{q} [\omega \hat{R}_{ij}^{(1)} + (1 - \omega) (\hat{X}_j^{(2)}/\hat{X}_j^{(1)}) \hat{R}_{ij}^{(1)}] = \hat{X}_j$$

and

$$\Sigma_{i=1}^{q} \hat{Z}_i = T.$$

This property, called the additivity, is often quite important in sample surveys where it is desirable to have the sum of estimates agree to fixed totals.

3. EFFICIENCY OF THE PROCEDURE

3.1 Variance of the estimates

To assess the efficiency of the procedure, we shall derive the variance of $\hat{Z}_1$ as a function of the variance of the simple estimator $\hat{Z}_1^{(1)}$ obtained

Assuming the relative variances of the $\hat{X}_j^{(1)}$'s are all equal to a constant "c", this expression then becomes, with the application of equation (2.5)

$$= [1 + c/(1 - \omega)] \; E \; Var \; \Sigma_{j=1}^{p} \hat{R}_{1j}^{(1)} \mid \hat{X}^{(1)}$$

$$= [1 + c/(1 - \omega)] \; E \; Var \; \hat{Z}_1^{(1)} \mid \hat{X}^{(1)} \qquad (3.2)$$

Finally, the application of equations (2.8) and (2.5) leads to

$$Cov \; \hat{Z}_1^{(1)}, \hat{Z}_1^{(2)} = Cov \; \Sigma_{k=1}^{p} \hat{R}_{1k}^{(1)}, \Sigma_{j=1}^{p} \hat{m}_{1j}^{(1)} \hat{X}_j^{(2)}$$

$$= \Sigma_{j=1}^{p} \; E \; Var \; \hat{R}_{1j}^{(1)} \mid \hat{X}^{(1)}$$

$$= E \; Var \; \Sigma_{j=1}^{p} \hat{R}_{1j}^{(1)} \mid \hat{X}^{(1)}$$

$$= E \; Var \; \hat{Z}_1^{(1)} \mid \hat{X}^{(1)} \qquad (3.3)$$

Thus from equations (3.1), (3.2) and (3.3)

$$Var \; \hat{Z}_1 = Var \; \hat{Z}_1^{(1)} - (1 - \omega) \; [Var \; E \; \hat{Z}_1^{(1)} \mid \hat{X}^{(1)} - c \; E \; Var \; \hat{Z}_1^{(1)} \mid \hat{X}^{(1)}] \; .$$

The variance expression becomes more explicit if we define index of uniformity (γ_1) of the characteristic C_1' in the categories $C_1, C_2, \ldots, C_p$

as
$$\gamma_1 = \frac{[E \; Var \; \hat{Z}_1^{(1)} \mid \hat{X}^{(1)}]}{Var \; \hat{Z}_1^{(1)}} = \frac{[Z_1 - \Sigma_{j=1}^{p} R_{1k} \; m_{1k}]}{(Z_1 - Z_1^2/T)}$$

where T is the total population size.

One may easily verify that $0 \leq \gamma_1 \leq 1$, that $\gamma_1 = 1$ if the characteristic C_1' is uniformly distributed in all categories $C_1, C_2, \ldots, C_p$ (in that case $m_{1k} = Z_1/T$ for all k) finally that $\gamma_1 = 0$ if the characteristic is

totally concentrated in some categories. The more uniformity there is, the larger is γ_1. Now the variance expression becomes

$$\text{Var } \hat{Z}_1 = [1 - (1 - \omega)\ (1 - \{1 + c\}\ \gamma_1)]\ \text{Var } \hat{Z}_1^{(1)} . \tag{3.4}$$

Consequently, $\hat{Z}_1$ is more efficient than $\hat{Z}_1^{(1)}$ whenever $(1 + c)\gamma_1$ is less than 1; this will be the case when a characteristic studied in the second phase is not uniformly distributed in a partition estimated in the first phase. The choice of the partition plays then an important role, as it is intuitively expected. On the other hand, the value of c should practically be relatively small since if the relative variance exceeds 6.25%, this would imply that the length of a 95% confidence interval would exceed the estimate itself. The gain in efficiency may be written as

$$G = (1 - \omega)\ [1 - (1 + c)\gamma_1] .$$

In the case where S is a complete census, it may be proved that

$$\text{Var } \hat{Z}_1 = (1 + c)\gamma_1 \ .\ \text{Var } \hat{Z}_1^{(1)} ; \tag{3.5}$$

as before, the proposed estimator is more efficient than $\hat{Z}_1^{(1)}$ whenever $(1 + c)\gamma_1$ is smaller than 1.

It is worth pointing out that this special estimation method considered here is currently used in the Canadian Labour Force Survey when making LFS data consistent with population projection estimates of 20 age-sex groups.

Let's consider that the interest is the estimation of the three LF characteristics "Employed" (E), "Unemployed" (U), and "Not in Labour Force" (N); the estimation procedure makes use of census information following the method described here. The matrix $\hat{M}^{(LFS)}$ is formed of elements $\hat{m}_{ij}^{(LFS)}$ (j = 1, 2, ..., 20; i = 1, 2, 3) equal to $\hat{R}_{ij}^{(LFS)}/\hat{X}_j^{(LFS)}$ where $\hat{R}_{ij}^{(LFS)}$ is the estimate from the LFS sample of the total number of persons in age-sex group "j" having the LF characteristic "i" and $\hat{X}_j^{(LFS)}$ is the

estimate from the LFS sample of the total number of persons in age-sex group "j". Then final estimates are obtained through the following matricial relationship:

$$\hat{Z} = \hat{M}^{(LFS)} X$$

More explicitly, the number of employed "E" is estimated by

$$\hat{E} = \Sigma_{j=1}^{20} (\hat{E}_j^{(LFS)}/\hat{X}_j^{(LFS)}) X_j$$

where $\hat{E}_j^{(LFS)}$ is the estimate of "employed" in age-sex group "j" from LFS sample

$\hat{X}_j^{(LFS)}$ is the LFS estimate of the number of persons in age-sex group "j"

and

X_j is the census population projection for age-sex group "j".

We may examine the efficiency of this method assuming SRS for LFS sample design and comparing with the variance of the original sample estimate $\hat{Z}^{(LFS)}$. We have established from (3.5)

$$\text{Var } \hat{E} = (1 + c)\gamma_E \text{ Var } \hat{E}^{(LFS)} .$$

Given the sample size of the LF survey, in the 20 age-sex groups, c is negligible. The index γ_E of the uniformity of the characteristic "employed" in the 20 age-sex categories as defined in LFS estimation procedure was calculated and is approximately equal to 0.67. The reduction in variance due to the application of the procedure is then of the order of 1/3. For the characteristic "unemployed", the reduction in variance is not significant, since the index of uniformity is close to 0.98. For the characteristic "not in labour force", it was found that γ_N = .63 resulting in a reduction in variance slightly larger than 1/3. In situations such as the LFS where c is negligible, substantial gain may be expected, especially when choosing properly the categorization of the census information.

3.2 Optimization

If we want to estimate each component of Z separately, we may consider the following estimate of say Z_1

$$\hat{Z}_1^* = r_1 \hat{Z}_1^{(1)} + (1 - r_1) \hat{Z}_1^{(2)}$$

where r_1 $(0 < r \leq 1)$ is a known constant. The optimum value of r_1 is given by

$$r_1 = \omega + (1 - \omega)\gamma_1 c/(1 - \gamma_1 + \gamma_1 c)$$

and the minimum variance is

$$\text{Var } \hat{Z}_1^* = [1 - (1 - \gamma_1)^2 (1 - \omega) (1 - \gamma_1 + \gamma_1 c)^{-1}] \text{ Var } \hat{Z}_1^{(1)} .$$

It may be observed that, with small values of γ_1 and c, the optimum choice of r_1 is close to ω and the gain in efficiency when using $\hat{Z}_1^*$ instead of $\hat{Z}_1$ is negligible; furthermore, this superiority of $\hat{Z}_1^*$ over $\hat{Z}_1$ no longer necessarily holds in practical situations where r_1 has to be estimated since, then, the minimum variance is not attained. One disadvantage of using $\hat{Z}_1^*$ instead of $\hat{Z}_1$ is that the additivity property no longer holds; indeed, there is no certainty than $\Sigma_{i=1}^{q} \hat{Z}_i^* = T$, the total population size.

4. APPLICATION TO COMPOSITE ESTIMATION

The estimation procedure developed here may be seen to have interesting properties when applied in the area of composite estimation. In periodical surveys such as the LF survey where the same basic information is collected monthly from a rotating sample, it is tempting to try to improve the estimates of the current month by incorporating somehow the information which was collected in preceding months, taking advantage of the high month-to-month correlation of the data. In the LF survey, households selected stay in the sample for six consecutive months; the rotation

pattern is such that each month, one-sixth of the LFS sample is renewed and each of the so-called rotation groups may be assumed to form an independent sample. Composite estimates may be constructed for a particular month, combining the information directly collected from the current sample with the information contained in the rotation samples which were dropped in the past months. A well known approach consists of taking a linear combination of the estimates derived for all months from all rotation samples, determining the parameters in such a way that unbiased (almost) estimates with minimum variance are obtained. The method proposed here is somewhat different; let us describe it first and then examine its properties.

Let us say that the rotation period is equal to K months and that K^{-1} of the sample is replaced each month. For the sake of simplicity, let us assume that the information collected in month "m", only, is used to improve the estimates of month "m+1". Let Z be the partition of the population to be estimated (say, as before, the total number of "Employed", the total number of "Unemployed" and the total number of "Not in labour force"). Then the estimates of Z may be presented in the following diagram for the different rotation samples:

	Sample S_1	Sample S_2	Sample S_3		Sample S_K	Sample S_1'
Month "m"	${}_m\hat{Z}^{(1)}$	${}_m\hat{Z}^{(2)}$	${}_m\hat{Z}^{(3)}$	...	${}_m\hat{Z}^{(K)}$	
Month "m+1"		${}_{m+1}\hat{Z}^{(2)}$	${}_{m+1}\hat{Z}^{(3)}$	...	${}_{m+1}\hat{Z}^{(K)}$	${}_{m+1}\hat{Z}^{(1)}$

In this diagram ${}_t\hat{Z}^{(r)}$ is the simple unbiased estimate of Z at time "t" from the rotation sample S_r. In month "m+1", the rotation sample S_1' has replaced the rotation sample S_1. Here again we shall assume the independence of the samples and the design SRSWR.

Our concern now is the improvement of the estimator ${}_{m+1}\hat{Z} = K^{-1} \Sigma_{r=1}^{K} {}_{m+1}\hat{Z}^{(r)}$ using the information available from month "m". Parallel to what was presented earlier, let us consider the following steps:

Step 1 - Using the information collected in both months from samples $S_2, S_3, \ldots, S_K$, let us estimate the transition matrix G (called the gross movement matrix) = (g_{ij}) where $g_{ij} = R_{ij}/{}_mZ_j$, R_{ij} being the number of persons of the population P having the characteristic C_j' in month "m", and the characteristic C_i' in month "m+1", and ${}_mZ_j$ being the number of persons of P having the characteristic C_j' in month "m". Let us call $\hat{G}$ this matrix.

Step 2 - Derive ${}^{(G)}_{m+1}\hat{Z}^{(1)}$ an estimator of Z in month m+1 from sample S_1 with the formula

$$ {}^{(G)}_{m+1}\hat{Z}^{(1)} = \hat{G}\ {}_m\hat{Z}^{(1)} $$

Step 3 - Combine ${}^{(G)}_{m+1}\hat{Z}^{(1)}$ with ${}_{m+1}\hat{Z}$ to obtain

$$ {}^{(G)}_{m+1}\hat{Z} = \omega \cdot {}_{m+1}\hat{Z} + (1 - \omega)\ {}^{(G)}_{m+1}\hat{Z}^{(1)} $$

In this particular situation the weight $\omega = K/(K+1)$ could be used. Under the given hypothesis and deriving similar proofs than the ones given earlier, it is possible to show that

i) ${}^{(G)}_{m+1}\hat{Z}$ is an unbiased estimate of Z for the month "m+1",

ii) ${}^{(G)}_{m+1}\hat{Z}$ preserves the additivity, and

iii) the variance of ${}^{(G)}_{m+1}\hat{Z}_i$ may be written in the following form (neglecting terms of the order of the relative variance of published estimates):

$$ \mathrm{Var}\ {}^{(G)}_{m+1}\hat{Z}_i = [1 + (K^2 - 1)^{-1} K \gamma_i] K/(K + 1) \cdot \mathrm{Var}_{m+1}\ \hat{Z}_i $$

where γ_i is, as defined earlier, the index of uniformity of the characteristic C_i' at the time "m+1" in the categories C_1', C_2', ..., C_q' at time "m"; for this specific application here, γ_i should be called the index of gross movement from month "m" to month "m+1" for the characteristic C_i'; this index will be "0" if the population is stable with respect to the characteristic C_i' and could theoretically reach the value "1" in the case of extreme mobility of the population with respect to C_i'.

For the sake of comparison, let us examine now a composite estimate in the form of a regression estimate.

$$ {}^{(R)}_{m+1}\hat{Z}_i = {}_{m+1}\hat{Z}_i + \lambda_i[({}_m\hat{Z}_i^{(1)} - \Sigma_{r=2}^{K} \; {}_m\hat{Z}_i^{(r)}/K - 1)] $$

with λ_i chosen to minimize the variances of ${}^{(R)}_{m+1}\hat{Z}_i$.

The optimum λ_i found is

$$ \lambda_i = (K - 1)\, \rho_i/K^2 \; [\mathrm{Var}\; {}_{m+1}\hat{Z}_i/\mathrm{Var}\; {}_m\hat{Z}_i]^{1/2} $$

where ρ_i is the correlation coefficient between ${}_m\hat{Z}^{(r)}$ and ${}_{m+1}\hat{Z}^{(r)}$.

The minimum variance expression is then

$$ \mathrm{Var}\; {}^{(R)}_{m+1}\hat{Z}_i = [1 - (K - 1)\, K^{-2}\, \rho_i^2]\; \mathrm{Var}\; {}_{m+1}\hat{Z}_i \; . $$

The mathematical expressions for ρ_i^2 and γ_i are respectively

$$ \rho_i^2 = (R_{ii} - {}_mZ_i\; {}_{m+1}Z_i/T)^2/{}_mZ_i\; (1 - {}_mZ_i/T)\cdot {}_{m+1}Z_i\; (1 - {}_{m+1}Z_i/T) $$

and

$$ \gamma_i = [{}_{m+1}Z_i - \Sigma_{j=1}^{q}\; (R_{ij})^2/{}_mZ_j]/{}_{m+1}Z_i\; (1 - {}_{m+1}Z_i/T) $$

with T being the total population size.

It follows then that

$$1 - \gamma_i \geq \rho_i^2 .$$

One may easily prove then, that whenever ρ_i^2 is not too small

$$\mathrm{Var}\ {}^{(G)}_{m+1}\hat{Z}_i \leq \mathrm{Var}\ {}^{(R)}_{m+1}\hat{Z}_i$$

A composite estimation procedure based on gross movement is then advantageously comparable from variance point of view to the one based on regression. Furthermore it has the two following properties:

i) The unbiasedness of the estimates is preserved; this is not rigorously true with the regression method when λ_i is estimated, and

ii) The additivity of the estimates is preserved: this property does not hold with the regression method since the optimum λ_i varies with the characteristic C_i' under study.

The values of γ_i and ρ_i^2 have been calculated using LFS data for the mutually exclusive categories "Employed", "Unemployed", "Kept house", "Attended school" and "Other". The percentages of reduction in sampling variance have been calculated for the simple case discussed here with the composite estimators ${}^{(R)}_{m+1}\hat{Z}_i$ and ${}^{(G)}_{m+1}\hat{Z}_i$ (using K = 6).

L.F. Classifications	Value of Parameters		% Reduction in Variance for	
	ρ_i^2	γ_i	$^{(R)}\hat{Z}_{i\,m+1}$	$^{(G)}\hat{Z}_{i\,m+1}$
Employed	.78	.21	10.8	11.2
Unemployed	.37	.62	5.1	5.2
Kept house	.87	.13	12.1	12.4
Attended school	.84	.15	11.7	12.0
Other	.78	.22	10.7	11.1

This Table shows that the difference in efficiency between the two estimators is not significant; the need for additivity of estimates becomes the important criterion then. Let us further mention that for K = 6 as it is in the LF survey, gains are not substantial for the simplified situation examined here when using only the previous month data. A more complete composite estimation procedure should take into account all available historical information. To this end, an iterative process could be developed leading to an important reduction in variances. Further work has to be conducted in this area.

The author wishes to acknowledge M.P. Singh and the referee for their suggestions leading to simplifications in the presentation.

RESUME

Cet article vise l'amélioration des estimations d'enquête par sondage dans le cas précis où l'échantillon d'enquête, ou une partie de celui-ci, est inclus dans un échantillon plus grand pour lequel de l'information auxiliaire est disponible. L'auteur y décrit certaines propriétés d'une méthode d'estimation - méthode d'ailleurs quelquefois employée dans certains cas particuliers - et établit les conditions de gain en efficacité. L'application de la méthode aux plans de sondage rotatifs pour enquêtes continues reçoit une attention spéciale dans le cadre de l'estimation composite.

REFERENCE

[1] Tenenbein, A., "A Double Sampling Scheme for Estimating from Misclassified Multinomial Data with Applications to Sampling Inspection", Technometrics, 1972, 187-202.

SOME VARIANCE ESTIMATORS FOR MULTISTAGE SAMPLING

G.B. Gray, M.A. Hidiroglou, M. Cairns
Household Surveys Development Division

J.N.K. Rao (1975) derived a general formula for estimating the variance in multistage sample designs. This general formula extends the previous results by Des Raj (1966) to the case where the conditional variance from a given primary sampling unit is a random variable. The authors reviewed Rao's paper for its application to Horvitz-Thompson and Yates-Grundy variance estimators as well as the variance estimator for the random group method by Rao, Hartley and Cochran (1962). The authors present an altered version of the Yates-Grundy variance estimators as a result of Rao's paper.

1. INTRODUCTION

Des Raj (1966) considered the general unbiased estimator of $Y = Y_1 + Y_2 + \ldots + Y_N$ given by

$$\hat{Y} = \sum_{i=1}^{N} a_{is}\, t_i, \tag{1.1}$$

where a_{is} $(i = 1, 2, \ldots, N)$ are real numbers pre-determined for every sample s with the restriction that $a_{is} = 0$ if $i \notin s$ and t_i is an unbiased estimator of Y_i. In order that the estimator be unbiased, the condition is that $E(a_{is}) = 1$ for every i. The variance of the estimator is then

$$V_s(\hat{Y}) = \sum_{i=1}^{N} Y_i^2\, V_s(a_{is}) + \sum_{i=1}^{N} \sum_{i \neq j} Y_i\, Y_j\, Cov_s(a_{is}, a_{js})$$

$$+ \sum_{i=1}^{N} \sigma_i^2\, E_s(a_{is}^2) \tag{1.2}$$

where E_s, V_s and Cov_s respectively denote expected value, variance and covariance over all possible samples s of first stage units. An unbiased estimate of $V_s(\hat{Y})$ is given by

$$v_s(\hat{Y}) = \sum_{i=1}^{N} b_{is}\, t_i^2 + \sum_{i=1}^{N} \sum_{i \neq j} d_{ijs}\, t_i\, t_j + \sum_{i=1}^{N} a_{is}\, \hat{\sigma}_i^2 \tag{1.3}$$

where like a_{is}, the real numbers b_{is}, d_{ijs} are predetermined real numbers such that $b_{is} = 0$ when $i \notin s$, $d_{ijs} = 0$ when i or $j \notin s$. On account of the unbiasedness property, we must also have that

$$E_s\, b_{is} = V_s(a_{is}) \text{ and } E_s\, d_{ijs} = Cov_s(a_{is}, a_{js}).$$

Here, σ_i^2 is the conditional variance within the ith unit (when selected) and the important point is that it does not depend upon the sample of first stage units drawn.

Rao (1975) extended Des Raj's formula to include cases where σ_i^2 depends upon the sample of first stage units drawn. Hence, the additional subscript s was added to σ_i^2 (σ_{is}^2) to take this point into consideration. An unbiased estimator of $V(\hat{Y})$ is then

$$v_s(\hat{Y}) = \sum_{i=1}^{N} b_{is}\, t_i^2 + \sum_{i=1}^{N} \sum_{i \neq j} d_{ijs}\, t_i\, t_j + \sum_{i=1}^{N} (a_{is}^2 - b_{is})\hat{\sigma}_{is}^2 . \tag{1.4}$$

Rao (1975) remarks that formula (1.4) is valid irrespective of the nature of σ_{is}^2. However, for the case $\sigma_{is}^2 = \sigma_i^2$, he suggests that formula (1.3) is preferable since it avoids the extra work of computing the b_{is}.

The authors applied the above development to particular sampling schemes, the most interesting one being the random group method.

2. APPLICATION TO HORVITZ-THOMPSON AND YATES-GRUNDY VARIANCE ESTIMATORS

In this section we discuss the variance estimates given by Yates-Grundy and Horvitz-Thompson and their relationship to Des Raj and Rao's generalized estimator.

A well known estimate of total given that the sampling of first stage units is without replacement and proportional to size is

$$\hat{Y} = \sum_{i=1}^{n} t_i/\pi_i \qquad \text{(Horvitz-Thompson)}$$

where π_i = inclusion probability of ith first-stage unit.

The variance and the estimate of variance are respectively

$$V(\hat{Y}) = \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1\right) Y_i^2 + \sum_{i=1}^{N} \sum_{i \neq j} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) Y_i Y_j + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i} \tag{2.1}$$

and

$$v(\hat{Y}) = \sum_{i=1}^{n} \frac{1}{\pi_i} \left(\frac{1}{\pi_i} - 1\right) t_i^2 + \sum_{i=1}^{n} \sum_{i \neq j} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) t_i t_j + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} \tag{2.2}$$

where π_{ij} = joint inclusion probability of ith and jth first-stage units.

These can be derived using Des Raj and Rao's generalized formulae for variance. In this case we have that

$$a_{is} = 1/\pi_i \quad \text{if } i \varepsilon s \text{ with } P(a_{is} = 1/\pi_i) = \pi_i$$

$$= 0 \quad \text{if } i \notin s \text{ with } P(a_{is} = 0) = 1 - \pi_i .$$

We have that $E(a_{is}) = 1$ for every i. Furthermore for every pair i, j whether equal or not,

$$\mathrm{Cov}_s(a_{is}, a_{js}) = E\, a_{is}\, a_{js} - E\, a_{is}\, E\, a_{js}$$

$$= \pi_{ij}/(\pi_i \pi_j) - 1 . \tag{2.3}$$

Since $\pi_{ij} = \pi_i$ when $i = j$, we have

$$V(a_{is}) = 1/\pi_i - 1 \tag{2.4}$$

and finally $E\, a_{is}^2 = 1/\pi_i$. (2.4a)

Substituting (2.3), (2.4) and (2.4a) into (1.2), we obtain (2.1) as required.

To derive the Horvitz-Thompson variance estimator using formula (1.3), we only need to obtain additional expressions b_{is} and d_{ijs}. Using equations (2.3) and (2.4), dividing $V(a_{is})$ by π_i and $Cov_s(a_{is}, a_{js})$ by π_{ij} we obtain that

$$b_{is} = 1/\pi_i \ (1/\pi_i - 1) \quad \text{for } i \varepsilon s$$

$$= 0 \quad \text{otherwise,}$$

$$d_{ijs} = 1/\pi_i\pi_j - 1/\pi_{ij} \quad \text{for } i, j \varepsilon s$$

$$= 0 \quad \text{otherwise,}$$

and

$$a_{is} = 1/\pi_i \quad \text{for } i \varepsilon s$$

$$= 0 \quad \text{otherwise.}$$

Considering formula (1.4) in light of the above parameters, we obtain that

$$a_{is}^2 - b_{is} = 1/\pi_i^2 - 1/\pi_i \ (1/\pi_i - 1)$$

$$= 1/\pi_i \quad \text{for } i \varepsilon s$$

$$= 0 \quad \text{otherwise.}$$

Hence, in the case of the Horvitz-Thomspon estimator, formulae (1.3) and (1.4) are identical.

Sen (1953), Yates and Grundy (1953) have expressed $V(\hat{Y})$ as

$$V_{YG}(\hat{Y}) = \sum_{i<j}^{N} (\pi_i\pi_j - \pi_{ij}) \ (Y_i/\pi_i - Y_j/\pi_j)^2 + \sum_i \sigma_i^2/\pi_i. \qquad (2.5)$$

It can be shown that upon expansion and re-arrangement of terms that for without replacement sampling schemes (2.5) is the same as (2.1).

The estimate of (2.5) as given by Sen (1953) and Yates-Grundy is

$$v_{YG}(\hat{Y}) = \sum_{i<j}^{n} (\pi_i\pi_j/\pi_{ij} - 1)\,(t_i/\pi_i - t_j/\pi_j)^2 + \sum_{i=1}^{n} \hat{\sigma}_i^2/\pi_i$$

$$= \sum_{i=1}^{n} t_i^2/\pi_i \sum_{j(\neq i)}^{n} (\pi_j/\pi_{ij} - 1/\pi_i)$$

$$+ \sum_{i=1}^{n} \sum_{j(\neq i)}^{n} t_i t_j [1/(\pi_i\pi_j) - 1/\pi_{ij}] + \sum_{i=1}^{n} \hat{\sigma}_i^2/\pi_i . \qquad (2.6)$$

The point to note in (2.6) is that the coefficient of t_i^2, b_{is}, is quite different from that in (2.2) while d_{ijs} is the same in both (2.2) and (2.6). Here,

$$b_{is} = [\sum_{j(\neq i)}^{n} \pi_j/\pi_{ij} - (n-1)/\pi_i]/\pi_i \neq \frac{1}{\pi_i}(\frac{1}{\pi_i} - 1) .$$

It is interesting to note that $E_s(b_{is}) = V_s(a_{is})$.

Rao (1975) pointed out that the two variance estimators given by (1.3) and (1.4) may not be identical. In the case of the Horvitz-Thompson variance estimates it was seen they are identical while in the case of the Yates-Grundy variance estimates they are not the same.

For the case of Yates-Grundy, $a_{is} = 1/\pi_i$ as before, while

$$a_{is}^2 - b_{is} = 1/\pi_i^2 - 1/\pi_i\,[\sum_{j(\neq i)}^{n} \pi_j/\pi_{ij} - (n-1)/\pi_i]$$

$$= n/\pi_i^2 - 1/\pi_i \sum_{j(\neq i)}^{n} \pi_j/\pi_{ij}$$

$$\neq 1/\pi_i .$$

Given that the ith unit has been selected in the sample, the conditional expectation of $a_{is}^2 - b_{is}$ is

$$E_s\,[(a_{is}^2 - b_{is})/i\varepsilon s] = n/\pi_i^2 - (1/\pi_i)\,E_s \sum_{j(\neq i)}^{n} \pi_j/\pi_{ij}$$

$$= n/\pi_i^2 - (1/\pi_i) \sum_{j(\neq i)}^{N} (\pi_j/\pi_{ij})\,(\pi_{ij}/\pi_i).$$

$$= 1/\pi_i, \text{ noting that } \sum_{j(\neq i)}^{n} \pi_{ij} = (n-1)\pi_i$$

Using equation (1.4), we obtain another form for the estimate of variance given by

$$v(\hat{Y}) = \sum_{i<j}^{n} (\pi_i\pi_j/\pi_{ij} - 1)\,(t_i/\pi_i - t_j/\pi_j)^2$$

$$+ \sum_{i=1}^{n} (n/\pi_i - \sum_{j(\neq i)}^{n} \pi_j/\pi_{ij})\hat{\sigma}_i^2/\pi_i$$

3. TWO-STAGE SAMPLING SCHEMES

a) Simple Random Sampling Without Replacement

To illustrate J.N.K. Rao's formula (1.4), take the case of two-stage sampling:

$$N \xrightarrow{\text{S.R.S.}} n$$

$$M_i \xrightarrow{\text{S.R.S.}} m_i$$

Here, an unbiased estimate of the population total $Y = \sum_{i=1}^{N} Y_i$ is given by

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} M_i\,\bar{y}_i.$$

where

$$\bar{y}_{i.} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}.$$

In terms of Rao's notation,

$$\hat{Y} = \sum_{u_i \varepsilon s} a_{is} \hat{Y}_i,$$

where

$$a_{is} = N/n \text{ if } i\varepsilon s \text{ and zero otherwise: here } \hat{Y}_i = M_i \bar{y}_{i.} .$$

For unistage cluster sampling,

$$f(\underset{\sim}{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \text{ is an unbiased}$$

estimator for $V(\hat{Y}_c)$ where

$$\hat{Y}_c = \sum_{u_i \varepsilon s} a_{is} Y_i \text{ and } Y_i = M_i \bar{Y}_{i.} .$$

In Rao's general notation,

$$f(\underset{\sim}{Y}) = \sum_{u_i \varepsilon s} b_{is} Y_i^2 + \sum_{u_i,} \sum_{\substack{u_t \varepsilon s \\ i<t}} d_{its} Y_i Y_t.$$

In our case, $b_{is} = \frac{N^2}{n} \left(\frac{1}{n} - \frac{1}{N} \right)$. By formula (1.4) an unbiased estimator of $V(Y)$ is given by

$$v(\hat{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \sum_{i=1}^{n} \left(M_i \bar{y}_i - \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{n} \right)^2$$

$$+ \sum_{u_i \varepsilon s} \frac{N}{n} M_i^2 s_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) .$$

This estimator agrees with the one given in Cochran [1, p. 304].

b) Random Groups Method

Another scheme used is that Rao-Hartley-Cochran's random group method is applied at the first stage and S.R.S. without replacement at the second stage

$$N \xrightarrow{\text{R.H.C.}} n$$

$$M_i \xrightarrow{\text{S.R.S.}} m_i$$

Then, the first stage estimator of total is

$$\hat{Y}_c = \sum_{i=1} \frac{P_i}{p_i} Y_i \quad \text{where} \quad P_i = \sum_{i^{th}\ \text{group}} p_t .$$

From Rao et al [1962], we have that

$$f(\underset{\sim}{Y}) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \left[\sum_{i=1}^{n} P_i \left(\frac{Y_i}{p_i} - \hat{Y}_c \right)^2 \right]$$

and hence,

$$b_{is} = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \left(\frac{P_i}{p_i^2} - \frac{P_i^2}{p_i^2} \right) .$$

An unbiased estimate of the population total given secondary units have been selected is:

$$\hat{Y} = \sum_{i=1}^{n} \frac{P_i}{p_i} M_i \bar{y}_{i.} \quad \text{where} \quad \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}.$$

$\hat{\sigma}^2_{is} = M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2$ is the estimate of variance for secondary units within primary unit i. Hence, an unbiased estimator of $V(\hat{Y})$ is given by

$$v_i(\hat{Y}) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2} \left[\sum_{i=1}^{n} p_i \frac{M_i \bar{y}_{i.}}{p_i} - \hat{Y} \right]^2$$

$$+ \sum_{i=1}^{n} \left[\frac{P_i^2}{p_i^2} - \frac{\Sigma N_i^2 - N}{N^2 - \Sigma N_i^2} \left(\frac{P_i}{p_i} - \frac{P_i^2}{p_i^2} \right) \right] M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 .$$

This estimator is different from the one given in Rao et al [1962]. The estimator in Rao et al has been derived by Des Raj [1966]. Hence, given an estimator, it is possible to construct more than one unbiased estimator of the variance of this estimator. Variances for these estimates of variance should be investigated: however, they may be too cumbersome to compute, for we would need to construct estimators for these estimates of variance. A possible solution would be to undertake a simulation.

The variance-estimator given by Rao et al [1962] and by Des Raj [1966] for the random group method may be derived by formula (1.3) and is valid only if m_i is fixed regardless of the group formation and set of 1st stage units selected. J.N.K. Rao however pointed out that in self-weighting samples m_i will depend upon the sample of 1st stage units so that one should really define $m_{is} = \delta M_i P_{is}/p_i$, where $P_{is} = \Sigma p_i$, which may vary from sample to sample, depending upon the groups so formed. In this latter case, where m_{is} does vary, there is no choice but to employ formula (1.4). In place of $M_i^2 \left(\frac{1}{m_{is}} - \frac{1}{M_i}\right) s_i^2$ one could employ the abbreviated symbol δ^2_{is} as an estimate of δ^2_{is} for any type of sample design within the ith unit.

ACKNOWLEDGMENT

The authors wish to thank the referee for some helpful suggestions.

RESUME

J.N.K. Rao (1975) a élaboré une formule générale pour évaluer la variance dans les plans d'échantillons à plusieurs degrés. Cette formule vient compléter les travaux de Des Raj (1966) lorsque la variance conditionnelle d'une unité primaire d'échantillonnage est une variable aléatoire. Les auteurs ont étudié l'application de cette formule aux estimateurs de la variance de Horvitz-Thompson et Yates-Grundy de même qu'à l'estimateur de la variance pour la méthode du groupe aléatoire de Rao, Hartley et Cochran (1962). Les auteurs présentent une version modifiée des estimateurs de la variance de Yates-Grundy à la suite du document de Rao.

REFERENCES

[1] Cochran, W.G., "Sampling Technique", 2nd edition, John Wiley & Sons.

[2] Connors, W.S., "An Exact Formula for the Probability that Two Specified Sampling Units Occur in a Sample Drawn with Equal Probabilities and Without Replacement", Journal of the American Statistical Association.

[3] Des Raj, "Some Remarks on a Simple Procedure of Sampling Without Replacement", Journal of the American Statistical Association, Vol. 61 (1966), pp. 391-396.

[4] Hartley, H.O. and Rao, J.N.K., "Sampling with Unequal Probabilities and Without Replacement", Annals of Mathematical Statistics, Vol. 33, No. 2, (1962), pp. 350-374.

[5] Hidiroglou, M.A. and Gray, G.B., "A Computer Algorithm for Joint Probabilities of Selection in Systematic PPS Sampling", Survey Methodology, Vol. 1, No. 1, (1975), pp. 99-108.

[6] Horvitz, D.G. and Thompson, D.J., "A Generalization of Sampling Without Replacement from a Finite Universe", Journal of the American Statistical Association, Vol. 47 (1952), pp. 663-685.

[7] Rao, J.N.K., Hartley, H.O., and Cochran, W.G., "On a Simple Procedure of Unequal Probability Sampling Without Replacement", Journal of the Royal Statistical Society, Series B, Vol. 27 (1962), pp. 482-491.

[8] Rao, J.N.K., "Unbiased Variance Estimation for Multi-stage Designs", Unpublished manuscript, May 1975.

[9] Sen, A.R., "On the Estimation of Variance in Sampling With Varying Probabilities", Journal of the Indian Society of Agricultural Statistics, Vol. 5 (1953), pp. 119-127.

[10] Yates, F. and Grundy, P.M., "Selection Without Replacement from Within Strata with Probability Proportional to Size", Journal of the Royal Statistical Society, Series B, Vol. 15 (1953), pp. 253-261.

THE METHODOLOGY OF THE CANADIAN TRAVEL SURVEY, 1971

A. Ashraf
Household Surveys Development Division

The Canadian Travel Survey, 1971 was the largest survey on travel of Canadian residents. This paper describes some important aspects of the methodology. Particular emphasis is given to the development of definitions in relation to the methodology, the sampling technique and interview strategy.

1. INTRODUCTION

The Canadian Travel Survey, 1971 (CTS 71) was the largest national household survey designed to collect data on domestic travel by Canadians. It was primarily developed, designed and managed by Statistics Canada on behalf of the Canadian Government Office of Tourism, Department of Industry, Trade and Commerce.

The travel industry, at least historically, has not been treated as an industry, partially due to its fragmentation and distribution throughout other more homogeneous industries such as transportation, accommodation, recreation and service industries. The statistics about these industries have been collected and analysed for their own purposes without specific references to their tourism components.

The travel industry in the past has been fulfilling its own needs for data on domestic travel by specialized and localized market research type surveys. The government agencies, for their part, have conducted surveys if and when warranted for their marketing needs. This piece-meal approach to the data of vital importance was less than satisfactory to the travel industry planners and government policy makers.

The governments, both Federal and Provincial, as well as the industry have commissioned and conducted a number of market research studies, but those, until very recently, were largely uncoordinated and uncomparable due to varying emphasis in objectives and concepts. There was, however,

an almost complete lack of base data on domestic travel on a national scale.

The Federal-Provincial Conference on Travel and Tourism in 1967 resolved to sponsor a travel survey at a national scale to provide the industry and governments with the much needed data on domestic travel for a more realistic policy formation about the industry.

2. OBJECTIVES

The primary objective of the Canadian Travel Survey 1971 was to provide the base data covering the whole calendar year 1971 on the volume and value of travel by Canadians in Canada in such depth and geographic breakdown as is practicable. This broad statement of the objective was spelled out in terms of volume, to mean, an estimate of the total number of Canadians that travelled during the calendar year, value to mean, an estimate of the total expenditure in dollars that incurred as a result of the above travels, depth to mean, further details of volume and value in meaningful components, and finally geographic breakdown to mean, the estimates to be produced would be at national levels, down to the level of geography where they could be useful. These levels would be Provinces and sub-provincial areas designated as Origin and Destination Zones.

The secondary objective of the CTS 71 was to develop suitable methodology and concepts in this relatively new context and to lay the foundation for future systematic development of travel and tourism statistics.

The objective of this paper is to present a comprehensive description of the CTS 71. A brief discussion of alternatives in arriving at the concepts for the subject matter and the sample design is included to provide a general background to the actual decisions. The description of various sampling stages, use of panels, origin-destination zones, etc. is dealt with in separate sections, in order to emphasize their particular functions in fulfilling the objectives of the CTS 71. The details are excluded to facilitate a conceptual understanding of the

methodology. However, they are contained in the Procedures and Interviewers Manuals.

3. CONCEPTS, DEFINITIONS AND CHARACTERISTICS

Travel is often described as a discretionary mobility of individual(s) in time and space. In this process, the individual(s) may or may not incur expenditure, prior to this departure (for the purposes of his departure), in transit and at his destination, should there be one. In the broadest sense, then travel has four basic dimensions, namely, individual, time, distance and currency. Each ingredient in itself can be described in further components. It therefore follows that if all components were measured to the last detail we would indeed have a very comprehensive knowledge about travel. In practice, however, this is seldom practicable. The demands of the policy makers, having generated the need for data with some general statements about objectives are rather elaborate, and this is where conceptual and definitional problems arise. The definitions have to be operationally feasible and the characteristics have to be readily identifiable. This conflict of interest, or perhaps the conceptual differences among the policy makers, economists and statisticians, is by no means unique to the CTS 71. It is in fact a perennial conflict between the policy makers, who are primarily concerned with the immediacy for information, having generated the need for data, and the economists and statisticians who are primarily concerned with the lucidity of concepts, feasibility of definitions and the meaningfulness of the data (particularly in the case of a new series) in relation to its analytical plans, and of course, ultimate objectives.

The concepts, definitions and characteristics that were employed in the design of the CTS 71 methodology evolved as a result of a series of discussions with the knowledgeable persons in the industry, government and the academic world in Canada and elsewhere, and experiment of these various concepts in the field before they were selected for the CTS 71. The Travel Research Planning Committee of the Federal-Provincial Conference on Tourism came out with "Standard Definitions and Classifications of Travellers and Traveller Accommodations" in October, 1971.

Three concepts that presented the most controversy were trip, stop and destination. The geographic diversity of Canada made the task of giving any uniformity to these concepts even more difficult. It was clearly understood that local travel was not important for this study, however, certain cities being larger in population and area than some provinces, there seemed to be a pull in opposite directions. Similarly, it was agreed that with certain exceptions, "stops" have a definite economic impact. It was argued that if there exists a reason for a stop, then efforts may be made to prolong the stop and thus enhance the regional equity. A destination on the other hand can be associated with specific areas in a large number of cases, but in certain isntances there in fact is no specific destination. The question was of a realistic designation of a destination to all such trips in such a way that it will be both practically feasible and analytically meaningful.

The determination of characteristic of the trips and population was another area where very little development had indeed taken place. It was obvious that in order to measure the demand side of the travel industry, characteristics should be chosen that would provide a meaningful measure.

It is thus that we arrive at our first constraint, namely, the extent of demand. Pragmatically speaking, for an undefined and ill described industry we should be seeking knowledge about the extremeties of the "industry" as we visualize it to be before seeking its various elements. The planners on the other hand, having had nothing to go on for so long, were understandably impatient with a methodical and thorough investigation which, to them, will be another period of uncertainty or indecision. However, at the federal level, accommodation of such desires from both public and private sector means trade-offs between highly competitive alternatives. The definition of trip was to embody a compromise. It is very difficult to arrive at the definition of trip without excluding certain type of trips. For example, an overnight trip to a nearby summer cottage, say 150 miles away. However, adopting a distance criteria may exclude the former and include the latter. On the other hand, a trip may be defined by time, e.g., all overnight trips. In this case, however,

we will be able to get both of the aforementioned trips, if, and only if, a night was spent at the cottage, otherwise both will be excluded, along with all other same day excursions, sight-seeing and business trips. One may at this point quite reasonably ask why not use both elements of the trip concept in defining it, e.g., say 100 miles away from home and/or overnight. This definition looks extremely good on the surface, since it will only exclude the trip to the nearby cottage if no night was spent there. However, it may include all other overnight lodgings that may have little relevance to travel.

Of the above three approaches, the distance criteria offers a greater merit, in as much as shorter trips would not be reported together with all non-travel related trips. The next question then is the distance criteria that should be nationally applicable. The political geography of Canada is quite challenging in this respect. The boundaries of Prince Edward Island and Ontario being what they are, how does one determine a minimum equitable trip distance? A '100 mile away from home' trip definition in P.E.I. would include much of the travel outside the province; a 25 mile trip definition in Ontario, on the other hand, may be a shopping trip in a place like Toronto. In this case P.E.I. would have virtually no data on inter-provincial travel, while Ontario may have a lot of data on inter-city travel, neither of which would be very satisfactory.

The optimal strategy would appear to employ a combination of time and distance criteria, such that the inclusion of less relevant trips and the exclusion of relevant trips is minimized. This strategy is then applied to all characteristics. It would be rather lengthy to write the deliberations of all the important characteristics. The definitions that were developed are given in Appendix I for reference.

4. SAMPLE DESIGN

The basic design is a stratified multistage replicated sample of households. The selected households were interviewed on four occasions during the one year period. The sampling frame excludes institutions, military

establishments, and other special areas such as hospitals, nursing homes, etc. Each province and territory is considered independently in the design, so as to allow for provincial estimates.

The sample size was approximately 13,000 households from a population of some 5.5 million households in 1971, a sample of 0.24 percent households. The allocation scheme excluded Prince Edward Island, Northwest and Yukon Territories, wherein the sample size was increased to allow for an estimated minimum reliability for the number of trips. The sample size was determined by the budgetary constraints.

The sample sizes by provinces and territories were as follows:

Newfoundland	288
Prince Edward Island	144
Nova Scotia	468
New brunswick	360
Quebec	3546
Ontario	4608
Manitoba	648
Saskatchewan	576
Alberta	936
British Columbia	1224
Yukon Territory	175
Northwest Territories	175
Canada Total	13148

A more detailed description of the design is in the following sections.

5. STRATIFICATION

The Canadian population is relatively speaking, densely populated in few urban areas, while the rest is scattered in rural areas. A random sample of households under such conditions is not possible due to two reasons. One, there does not exist an "Address Register" of all

dwellings/households that can be used as a sampling frame, and two, the characteristics of population in all areas of a province are not necessarily homogeneous. It was therefore deemed desirable to consider a stratified, multistage design. The stratified form allows us to group together those areas of population that are homogeneous with respect to the characteristics that are related to the objectives of the study. The total area can then be divided into a given number of sub-sets that are initially exclusive and exhaustive, and an independent sampling scheme may be devised for each stratum.

In the CTS 71, therefore, stratification was one of the design features. The stratification variables were not all readily available. The stratification variables should be related to the travel habits of the population, and/or in part, the populations propensity to travel. Whatever the size of the stratum, a fairly large portion of the population in that region should have this characteristic. The travel as discussed in the preceding section may be for any reason associated with business or leisure. It may then be assumed that the population in the vicinity of business locations may be prone to more frequent business travel than away from it. This is not to say that population away from business areas would not travel for business purposes, but their travel would be less frequent.

On the basis of the above considerations, each province was stratified into two types of areas: Large Urban and Non-large Urban. This was the first stage of stratification, and therefore a given province now had one or morelarge Urban and Non-large Urban strata. The large areas comprising LUs and NLUs were further stratified into a number of other strata of a more manageable number and size for practical as well as design reasons. The practical reasons were, of course, the cost of various field operations, whereas the design reasons were the variances within and between strata and stages of sampling. Again, due to non-availability of information on travel related characteristics, the development of a cost and variance model was not possible. The factors that were considered instead were geography, population density, cost of listing, enumeration, etc.

A number of strata were constructed in each province comprised of contiguous Enumeration Areas. An EA is a Census Enumerator's Assignment and consists of approximately 150-250 households. The geographical boundaries of an EA respects all other physical and political boundaries. The number of EAs per stratum and the size of each stratum was decided on the basis of factors such as the overall sample allocation for NLU and LU areas and the expected number of households per EA.

The demographic data available at the time of the design was from the 1966 Census. The change in the growth and structure of the population by 1971 was inevitable. The exact magnitude of the change could not be measured at the time. The strata were therefore constructed of equal sizes, and the sample size for each stratum was approximately equal.

Large Urban Areas

The LUs were urban centres such as St. John's in Newfoundland, Regina in Saskatchewan, etc. A complete list of all LUs is given in Table 2. Within each LU there was sufficient heterogeneity in the socio-economic structure of the population to warrant further stratification. Guidance was derived from Census of Population designated areas such as Census Tract which are areas of socio-economic uniformity. Large Urban areas were further stratified to account for economic variations that may be pertinent to the area. Income and size of the family were two criteria that were used for this purpose. In certain LUs the second stage of stratification entailed designating areas of very high income. The contiguous Census Tracts with average household income of $8,000 or more were grouped to construct these strata. This stratification was carried out in those areas where such areas were large enough in proportion to the rest of the area to permit creation of such stratum, otherwise such desparities in population characteristics were ignored. The areas were instead stratified into strata of equal size of contiguous Census Tracts that registered differences. For example, a city such as Montreal had 21 strata in all, 2 of very high income households, 2 of Apartment Buildings and the 17 were all other households areas.

Non-Large Urban Areas

The NLU areas, were the areas other than the LU areas. The stratification of NLU areas was done on the basis of two factors, cost and similarity of characteristics in the contiguous EAs in the population. The basic unit was an EA which was also the first stage sampling unit. The number of strata for each province was arrived at by considering the sample size, expected number of EAs per stratum and the enumerator assignment.

6. ORIGIN-STOP-DESTINATION ZONES

The Origin-Stop-Destination Zones are sub-provincial areas or more specifically a grouping of strata for which tabulation was required. Therefore the delineation of the Origin-Stop-Destination Zones (O-D) zones was carried out in conjunction with the stratification. It may not appear to be a design problem, however, from the tabulation requirements that were made at the time for the various geographic levels, the O-D zones were indeed to conform with the strata boundaries. As noted in the discussion on concepts and definitions, a trip includes areas where stops are made and one of those stops may be the destination of the trip. It is where the interview took place. The stops, including the destination may be anywhere. Clearly some level of aggregation was warranted here. The Tourist Industry and Provincial governments of course had Tourist Regions in mind. These Tourist Regions were quite numerous and furthermore, their boundaries were not particularly related to any other boundaries. Considering the sample size and the assumption of average number of trips by a household, a maximum of 35 O-D zones was agreed upon. The 35 zones were arrived at by considering factors such as population homogeneity and density, geographic contiguity, tourist attraction areas as defined by provinces and the overall conformity of these areas to a number of strata. This allowed for the summation of characteristics over a given number of strata that correspond to an O-D zone.

The above factors were not necessarily applied to each province, but any one or more of them were used as a criteria. For example in British

Columbia, Vancouver (p. 892,286) and Victoria (p. 300,000) is collapsed to one O-D due to their highly urbanized nature in relation to the rest of the province. The northern parts form another zone due to the population engagement in ranching or forest industry. Aimilarly, that portion of British Columbia known as the Fraser Valley (excl. Victoria) were combined to form an O-D zone as the main industries in these parts are dairy farming, market farming and fishing.

For each province, distinction between LU and NLU has been maintained whenever possible. For example in Ontario, Windsor, London, Kitchener-Waterloo areas (p. 612,368) are combined to form an O-D zone while Toronto (p. 2,158,496), Ottawa (p. 384,397) and Sudbury (p. 117,075) each form O-D zones by themselves.

In the above design of O-D areas, each area by definition is an origin, stop or destination with the exception of Rocky Mountain areas which are Stop/Destination areas, only. The trips originating from this area are combined with the usual O-D zone. Similarly Hull, Quebec, is a special case so that statistics on the National Capital Region may be produced. The distribution of Origin-Stop and Destination areas by province was as follows:

Newfoundland	2	
Prince Edward Island	1	
Nova Scotia	2	
New Brunswick	2	
Quebec	7	
Ontario	9	
Manitoba	2	
Saskatchewan	2	
Alberta	3	(2 O-D plus 1 Destination only)
British Columbia	5	
Canada	35	

7. SAMPLE SELECTION METHODS

The basic design was the multistage stratified replicated sample. The multistage design, as the name implies, is the division of a population in groups that are in a descending order of density. A city, for example, may be considered in two groups, i.e. localities and/or blocks within localities. A sample may then be selected of localities and blocks, rather than of blocks that may well be all over the city or clustered in one area.

In the NLU areas, the first stage unit was an EA, the second stage was the household, and the third and final stage was a person within the household. In LU areas, the census tracts were the first stage, blocks the second, household the third, and the person the fourth. Within LU areas, wherever the apartment building necessitated a stratum, the apartment building was the first stage, household the second, and the person the third and final stage.

First Stage Units

The first stage units were selected with probability proportional to size in all strata. In the LU non-apartment strata, two census tracts were selected with PPS (Census 1966), whereas in apartment strata, the total number of apartments to be selected in each stratum were divided into two replicates of equal sizes. The measure of size in this case was the number of apartments in each apartment building. Each of the two replicates were then selected independently with PPS. In the NLU, two equal number of replicates of EAs were selected with PPS. The measure of size was the population as of Census 1966. There were some strata where the number of EAs to be selected were not divisible by 2. In those cases, two strata were used to balance the number of EAs to be selected. For example, if two strata required the selection of 18 EAs or 9 EAs each, the 10 EAs were selected from one stratum and 8 EAs from the other, providing us with two replicates of equal sizes each, 4 and 5 in this case.

Second Stage Units

In the LUs where the second stage units were the city blocks, the selection procedure was the same as described above for the selection of first stage units for EAs and apartment buildings. That is, two independent replicates of 8 blocks were selected with PPS from each stratum. At this stage, the sample of city blocks, EAs and apartment buildings were delineated on the maps for the listing of dwellings in each one of them. All such maps were then sent to the appropriate Regional Offices for that purpose. The listing operation provided the more recent count of dwellings in each selected area. The sampling ratio were then calculated based on the recent listing and the expected number of households to be selected for that area. In each case a systematic random sample of households was selected.

Third Stage Units

The third stage unit was the person within a household, except in non-apartment LU strata, where this would be the fourth stage unit. The procedure in each case was the same. One person, 14 years of age and over was selected randomly from each household. The list of selected households was provided to each enumerator, prior to the date for the first interview. The household information was designed to be recorded on a household card. The household composition was required to be listed in two sections. One section required the listing of all household members that were aged 14 years or more as of the date of interview, in the descending order of age. Each entry on this section, therefore, was numbered from 1 onwards. A table of random numbers was used to determine which number, i.e. person is then selected. The person thus selected would then be interviewed for the trips that he took with or without other members of the household. The other section of the household information card would list all persons under 14 years of age in the household.

The sample design can thus be summarized as follows:

Table 1: Sample Design

Stage of Sampling \ Stratum	Large Urban Areas			Non-Large Urban Areas
	Apartment Buildings	Residential Homes	Residential Homes (Very high income)	Congiguous EAs
First Stage	Apartment Building	Census Tract	Census Tract	EA
Second Stage	Apartments	City blocks	City blocks	Households
Third Stage	Persons	Households	Households	Persons
Fourth Stage	None	Persons	Persons	None

8. INTERVIEW METHODS

The CTS 71 was designed to collect data on travel for the calendar year 1971. In the pilot study prior to the main CTS in 1971 it was determined that depending on the method of aid to recall, the optimum recall period lies somewhere between 12 and 16 weeks. The interview scheme, therefore had three basic features, personal interview, staggered over a year so as to be approximately 12 to 16 weeks apart and an aid to recall left with the respondent. The aid to recall selected for the CTS 71 was a pocket diary, designed so that essential information about the travel may be noted. The essentials of a trip thus recorded, it would then be not only more convenient to convey such trip information to the enumerator at the time of the interview but also less susceptible to loss of information. In order to stagger the length of recall, the entire sample size was randomly divided into three "panels" of equal sizes. The time between the interviews of each panel was distributed so as to allows different lengths of

recall for each of the 4 rounds of interview. The interviews were scheduled for the spring, summer and fall of 1971 and early (January) 1972. The interview scheme was as follows:

Table 2: Interview Scheme

Round	Panel	Interview Week
First	1 2 3	March 8-12 March 29-April 2 April 26-April 30
Second	1 2 3	June 7-11 July 5-9 August 2-6
Third	1 2 3	September 6-10 October 4-8 November 1-5
Fourth	1 2 3	January 10-14 January 17-21 January 24-28

The concept of bounded recall was introduced by way of New Years Day. Plans called for leaving the diary with the respondent at the beginning of the year. The operational problems, however, did not allow this to happen until the second round of interviews. An additional week was allowed to follow up the cases wherever the respondent was temporarily absent, but could be reached a few days later.

9. ESTIMATION

The method of estimation that follows is a straightforward application of multistage stratified sampling. The estimates incorporate the adjustment of weights in two steps. Firstly, the weights are adjusted at the

stratum level to account for non-response. The assumption here is that the characteristics of the respondent are similar to those of non-respondents. Secondly, the provincial population is estimated from the sample and then compared with the known population for 1971. The weights are then adjusted by the ratio of the known population to the estimated. The weights thus adjusted are used to estimate characteristics and the estimate of the variance of the estimated characteristics as shown in the following development.

A - Notations

W = Weight
a = Census Tract
b = Block, Apartment Building in Enumeration Area
c = Household
d = Person
t = Trip
e = Province
f = Stratum

In each stratum two replicates of equal sizes were selected. Estimate of a characteristic for a stratum is obtained by taking a mean of the estimates from the replicates. In the estimation procedure described below, the notation for replicates and various steps leading to the estimates are omitted for the sake of brevity.

B - Weights

In accordance with the design, the weight for each household is defined as the product of weights of different sampling stages. In order to better understanding of the estimation procedure, reference should be made to section 7 where selection methods are described:

i) In the Large Urban areas, then the weight for a household in High and Moderate income sub strata is defined as:

$$W_{ef} = W_a \cdot W_b \cdot W_c \cdot W_t$$

where

W_t = $\frac{\text{Total number of persons in the household}}{\text{Total number of persons on the trip}}$

W_d = Total number of persons in the household over 14 years of age.

W_c = Inverse of the probability of selection of (the interviewed) households within a block.

W_b = Inverse of the probability of selection of blocks within a Census Tract.

W_a = Inverse of the probability of selection of Census Tracts within a stratum.

The weights for the Apartment Building sub-strata are defined in much the same way as above except there are only three stages we therefore have

$$W_{ef} = W_b \cdot W_c \cdot W_t$$

where W_b = Inverse of the probability of the selection of Apartment Buildings.

ii) In the Non-Large Urban areas the weights are defined as

$$W_{ef} = W_b \cdot W_c \cdot W_t$$

where W_b = Inverse of the probability of the selection of Enumeration Areas.

C - Adjustment of Weights

The weights are adjusted for two factors: first for non-response and second for the "known" population totals.

The adjustment for non-response is made by assuming that the interviewed households have characteristics similar to those that cannot be interviewed. The adjustment is made at the "block" level by the following method:

Let W'_c = Inverse of the probability of the selection of households

s = Number of households selected in a block

s' = Number of households interviewed in a block

s^* = Number of households vacant in a block.

We then define W_c, the adjusted household weight

$$W_c = W'_c \left(\frac{s - s^*}{s'} \right)$$

The adjustment for "known" population totals is made at the provincial level. The adjustment factor is the ratio of the Census population projection and the estimate of population from the CTS.

Let P_e = The Population Projection for 1971 from the Census for province e

$\hat{P}_e$ = The estimate of population from the CTS 71 for province e

γ_e = The Adjustment factor for "known" population for the province e.

where

$$\gamma_e = \frac{P_e}{\hat{P}_e}$$

Then W^*_{ef}, the weight adjusted for "known" population total is given by

$$W^*_{ef} = W_{ef} \cdot \gamma_e$$

D - Estimates

The estimate $\hat{X}_e$ of a characteristic is therefore defined as

$$\hat{X} = \sum_f W^*_{ef} x_{ef}$$

$$= \frac{1}{2} \left(\sum_f {}^1W^*_{ef} \, {}^1x_{ef} + \sum_f {}^2W^*_{ef} \, {}^2x_{ef} \right)$$

where the superscripts 1 and 2 denote the replicate 1 and 2 respectively.

The estimate of variance, $\hat{V}$ of $\hat{X}$ is obtained as follows $\hat{V}(\hat{X}) = \frac{1}{4} ({}^1\hat{X}_{ef} - {}^2\hat{X}_{ef})^2$

10. ACKNOWLEDGEMENT

Acknowledgement is due to Mr. R. Platek and Mr. P.F. Timmons who directed the development and implementation of the methodology of the Canadian Travel Survey 1971, to Dr. M.P. Singh for his invaluable suggestions in improving the quality of this manuscript and finally to the referee for a very thorough and constructive criticism.

Appendix I - Definitions

Trip: 1) Travel to a plate 100 miles or more away from home excluding --
2) Travel to a place 25 miles or more away from community boundary

Destination: A place farthest away from the origin of the trip.

Stop: A place between origin and destination of a trip where either a night was spent or an expenditure was made of $1.50 or more, per person on the trip.

Time Period: Calendar year 1971. This included all the trips terminating in the year 1971.

Distance: Anywhere.

Expenditure: All expenditure in relation to the trip that would not have been incurred had the trip not taken place. This included expenditure at origin in preparation of trip, on the trip at stops and destination, on transportation, accommodation, sightseeing, entertainment, gifts, local transportation, etc.

Population: Canadian residents in the calendar year 1971 excluding --

Population Characteristics: Age, sex, marital status, income, occupation, education, type of employer, automobile ownership, etc.

RESUME

L'enquête sur les voyages des Canadiens pour 1971 a été la plus importante enquête effectuée dans ce domaine. Le présent exposé décrit quelques-uns des principaux aspects de la méthodologie. On insiste plus particulièrement sur l'élaboration de définition en fonction de la méthodologie, la méthode d'échantillonnage et les techniques d'interview.

METHODS TEST PANEL PHASE II - DATA ANALYSIS

R. Tessier
Household Surveys Development Division

In the Methods Test Panel Phase II it was required to do analysis of variance on proportions. Since such analysis gives only approximate results, two models were used in order to be able to draw safe conclusions. Analysis of variance was performed with the proportions as variable and also with the arc sine of the square root of the proportions. The two models are outlined in the present paper and empirical comparisons are made using the MTP Phase II data.

1. INTRODUCTION

The Methods Test Panel (MTP) Phase II study was an extensive study of the impact on the Labour Force Survey (LFS) data of extra burden imposed on the respondents. Extra burden was in the form of two supplementary questionnaires to the LF questionnaire.

Since many aspects need to be considered in such a study (eg. effects of burden itself, effect of procedure used to present the extra burden, effect of rotation groups), the resulting model for analysis becomes quite complex. It was, therefore, decided that an analysis of variance model would be most appropriate.

The data analysed in the study, though, are all in the form of proportions. Therefore, it cannot be readily assumed that they are normally distributed variables with the same variance. To see the impact of this approximation, two analysis were done concurrently, one using proportions and the other using the arc sine transformation of the square root of the proportions. Though other types of transformations could have been considered (see [1], [2]) it was decided, due to time constraints, to restrict the analysis to these two models.

2. THE DESIGN

The study consisted in presenting to the respondents, in addition to the

LF questionnaire, two supplementary questionnaires, the Consumers Finance questionnaire and the Household Facilities and Equipment questionnaire. The sample was divided into quarters and respondents in each quarter were approached differently. The four approaches are:

- One quarter of the respondents received one supplementary questionnaire in a month and the other in the following month, the enumerators using the regular LF interviewing procedure (called method A).

- One quarter of the respondents received both supplementary questionnaires in the same month, the enumerators using the regular LF interviewing procedure (called method B).

- One quarter of the respondents received one supplementary questionnaire in a month and the other in the following month, the enumerators using a special interviewing procedure (called method C).

- One quarter of the respondents received both supplementary questionnaires in the same month, the enumerators using a special interviewing procedure (called method D).

Further, a few months later, the Job Mobility supplementary questionnaire was presented to all four groups in order to study long term effects.

Five rotation groups were also used as an additional factor in the analysis in order to determine possible effect due to their use. Notice that there are six rotation groups on the regular LFS, but it was thought to be sufficient to use five of them for the test. This resulted in a three-way layout as presented in Table 1. For complete details on this study see [4] and [5].

Table 1: Three-way Layout for Respondent Burden Study

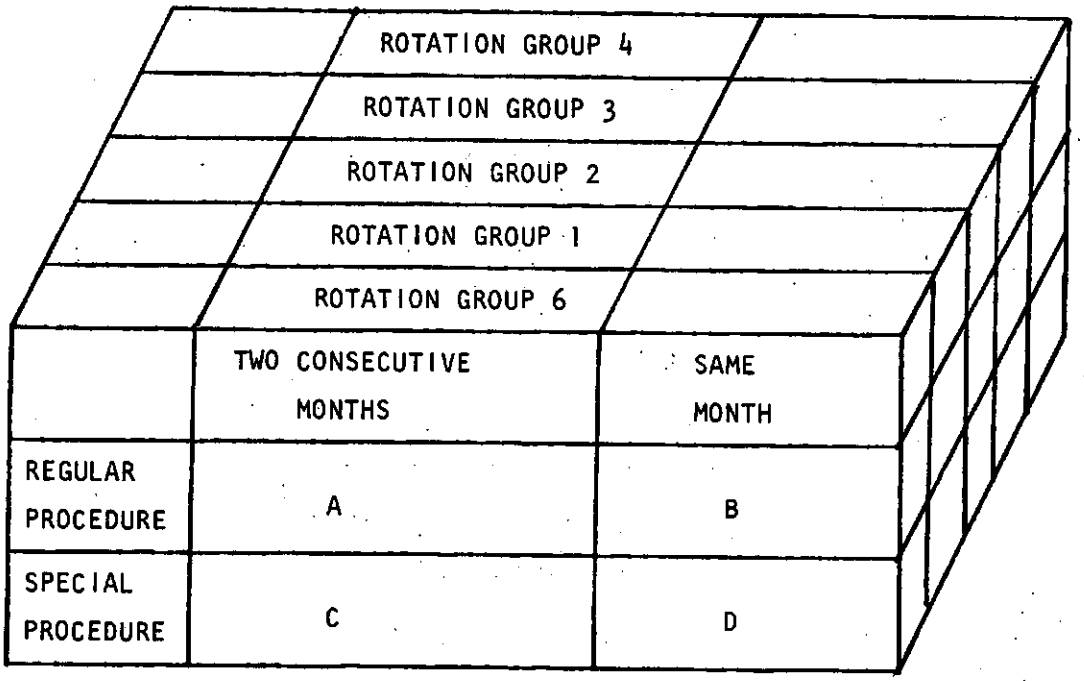


Table 2 provides the expected number of households for each cell of the three-way layout.

Table 2: Expected Number of Households

Rotation Group No. \ Method	A	B	C	D	All
1	66	64	80	78	288
2	82	73	69	78	302
3	74	92	63	77	306
4	71	68	85	63	287
6	77	73	71	78	299
All	370	370	368	374	1482

3. STATISTICAL ANALYSIS

Though a three-way layout analysis of variance model would have been appropriate, it was decided, for simplicity, to reduce the model to a two-way layout and adjust the tests by using corresponding tests on contrasts. Furthermore, since only one value is available per cell (proportion in the cell) and interactions were not assumed all equal to zero, it was impossible to estimate the error sum of squares (MSE). To remedy this situation the households of each cell were randomly split into two groups and proportions were calculated for each group, thus providing two values per cell and therefore permitting estimation of the error sum of squares. The resulting two-way layout model, where procedure and burden are combined, is presented in Table 3, where we define p_{ijk} as the proportion (having the characteristic) of units having received procedure and burden i, in rotation group j and randomly allocated to half cell k.

Table 3: Two-way Layout (Two Observations per Cell)

Procedure	Burden \ Rotation Group	1	2	3	4	6
Regular Procedure	Two Consecutive Months	p_{111}, p_{112}	p_{121}, p_{122}	p_{131}, p_{132}	p_{141}, p_{142}	p_{151}, p_{152}
	One Month	p_{211}, p_{212}	p_{221}, p_{222}	p_{231}, p_{232}	p_{241}, p_{242}	p_{251}, p_{252}
Special Procedure	Two Consecutive Months	p_{311}, p_{312}	p_{321}, p_{322}	p_{331}, p_{332}	p_{341}, p_{342}	p_{351}, p_{352}
	One Month	p_{411}, p_{412}	p_{421}, p_{422}	p_{431}, p_{432}	p_{441}, p_{442}	p_{451}, p_{452}

In sub-section 3.1 tests are derived using the proportions as variable, assuming that the proportions are normally distributed with variance equal to σ^2/n_{ijk} and sub-section 3.2 presents tests when the arc sine transformation is applied to the square root of the proportions.

3.1 ANOVA with proportions

Usually in analysis of variance with one observation per cell interactions are all assumed to be equal to zero (that is, the model is assumed to be

additive) since there are not enough degrees of freedom to test for interaction. But since in this case it was possible to randomly split the cells in two halves and obtain two observations per cell it was decided to test the hypothesis of additivity of effects. Therefore, the underlying assumptions are

$$\Omega: \begin{cases} p_{ijk} = \nu_{ij} + e_{ijk} \\ \{eijk\} \text{ are independent} \quad N(0,\sigma^2/n_{ijk}) \end{cases}$$

and the hypothesis to be tested is

$$H_o: \begin{cases} p_{ijk} = \alpha_i + \beta_j + e_{ijk} \\ \{eijk\} \text{ are independent} \quad N(0,\sigma^2/n_{ijk}), \end{cases}$$

where α_i refers to the procedure x burden effect and β_j to the rotation group effect.

Transforming e_{ijk}'s to obtain variables with constant variance per cell (σ^2) leads us to consider the following sums of squares for testing H_o:

$$\sum_{ijk} n_{ijk} (p_{ijk} - \nu_{ij})^2 \tag{3.1}$$

and

$$\sum_{ijk} n_{ijk} (p_{ijk} - \alpha_i - \beta_j)^2, \tag{3.2}$$

where n_{ijk} is the number of units in k^{th} half cell (i, j). Adjustment of variables by the number of units in half cells was required since the n_{ijk}'s were far from being constant from one cell to another.

The least-squares estimate of ν_{ij} found by minimising (3.1) with respect to ν_{ij} is

$$\hat{\nu}_{ij} = \frac{\sum_k n_{ijk} p_{ijk}}{\sum_k n_{ijk}} .$$

Minimizing (3.2) with respect to α_i and β_j does not lead to simple least-squares estimates of the parameters; indeed, if we minimize (3.2) with respect to α_i and β_j and eliminate β_j in the equations for α_i we obtain the following system of equations

$$Q_i = \sum_j \left(\sum_k n_{ijk}\, p_{ijk} - \frac{n_{ij.}}{n_{.j.}} \sum_i \sum_k n_{ijk}\, p_{ijk}\right) = \sum_{i'} C_{ii'}\, \alpha_{i'}, \tag{3.3}$$

where

$$C_{ii} = n_{i..} - \sum_j \frac{n_{ij.}^2}{n_{.j.}} ,$$

$$C_{ii'} = -\sum_j \frac{n_{ij.}\, n_{i'j.}}{n_{.j.}} , \quad i \neq i' \tag{3.4}$$

and where a dot represents summation over the index. In order to solve this system of equation the best side condition is $\alpha_4 = 0$. Also, an estimate of β_j is not required since by replacing β_j by its appropriate expression in terms of p_{ijk} and α_i in the differenciation of (3.2) with respect to β_j, we find that

$$SS_o = \sum_{ijk} n_{ijk} \left(p_{ijk} - \sum_{ik} \frac{n_{ijk}\, p_{ijk}}{n_{.j.}}\right)^2 - \sum_{ij} n_{ij.} \left(\hat{\alpha}_i - \sum_i \frac{n_{ij.}\, \hat{\alpha}_i}{n_{.j.}}\right)^2.$$

We, therefore, have the following F statistic to test for additivity of effects

$$F = \frac{SS_o - SS_\Omega}{SS_\Omega} \; \frac{IJ(K-1)}{(I-1)(J-1)}$$

which has an F distribution with $(I-1)(J-1)$ and $IJ(K-1)$ degrees of freedom, where

$$SS_o - SS_\Omega = \sum_{ij} n_{ij.} \left[\sum_k \frac{n_{ijk}\, p_{ijk}}{n_{ij.}} - \sum_{ik} \frac{n_{ijk}\, p_{ijk}}{n_{.j.}}\right]^2$$

$$- \sum_{ij} n_{ij.} \left(\hat{\alpha}_i - \sum_i \frac{n_{ij.}\, \hat{\alpha}_i}{n_{.j.}}\right)^2 ,$$

$I = 4$ (procedure x burden),
$J = 5$ (rotation groups),
and $K = 2$ (number of half cells).

If the hypothesis H_o is true, then, we test the following three contrasts:

$$\psi_1 = \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 , \tag{3.5}$$

$$\psi_2 = \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 , \tag{3.6}$$

$$\psi_3 = \alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 , \tag{3.7}$$

the hypothesis being

H_1: $\psi_1 = 0$, or no effect due to the special procedure,

H_2: $\psi_2 = 0$, or no effect if we ask the two supplementary questionnaires in the same month or in two consecutive months,

H_3: $\psi_3 = 0$, or no interaction between the procedures and the months in which the supplementary questionnaires are asked,

with the underlying assumptions being, now, H_o.

In order to find a simple expression for the variance of the contrasts, we first express them in terms of vectors; therefore, let us define

$$\underline{h}_1' = (1, 1, -1) , \underline{h}_2' = (1, -1, 1) , \underline{h}_3' = (1, -1, -1), \tag{3.8}$$

and $\underline{\alpha}' = (\alpha_1, \alpha_2, \alpha_3)$;

then $\psi_\ell = \underline{h}_\ell' \underline{\alpha} , \quad \ell = 1, 3$

since we have set $\alpha_4 = 0$ as the side condition. Furthermore, let us define

$$\underline{Q}' = (Q_1, Q_2, Q_3) \tag{3.9}$$

and the matrix $C = (C_{ij}), \quad i = 1, 3$
$j = 1, 3$

where Q_i and C_{ij} are defined in equation (3.3) and (3.4) respectively. We have now that the estimate of the contrasts can be expressed as

$$\hat{\psi}_\ell = \underline{h}_\ell' \, C^{-1} \, \underline{Q}, \quad \ell = 1, 3$$

and
$$V(\hat{\psi}_\ell) = \underline{h}_\ell' \, C^{-1} \, \Sigma_{\underline{Q}} \, C^{-1} \, \underline{h}_\ell, \; \ell = 1, 3$$

where $\Sigma_{\underline{Q}}$ is the variance-covariance matrix of the vector $\underline{Q}$ and since the matrix C is symmetric. The variance expression reduces to

$$V(\hat{\psi}_\ell) = \underline{h}_\ell' \, C^{-1} \, \underline{h}_\ell \, \sigma^2$$

since we have that

$$\Sigma_{\underline{Q}} = C \, \sigma^2$$

and the matrix C is of rank 3 and non-singular.

To find an estimate of σ^2 we use the data at the half cell level, that is, we find the error sum of squares as

$$SSE = \sum_{ijk} n_{ijk} \left(p_{ijk} - \sum_k \frac{n_{ijk} \, p_{ijk}}{n_{ij.}}\right)^2$$

and therefore,

$$\hat{\sigma}^2 = MSE = \frac{SSE}{IJ(K-1)} \; .$$

Thus, for testing $\psi_\ell = 0$ (ℓ=1, 3) we have the following statistics:

$$F_\ell = \frac{\psi_\ell^2}{\underline{h}_\ell' \, C^{-1} \, \underline{h}_\ell \, MSE}, \qquad \ell = 1, 3$$

which are distributed as F's with 1 and IJ(K-1) degrees of freedom.

3.2 ANOVA with a Transformation of the Proportions

In the ANOVA model using proportions as the variable, we have made the assumption that the variance of the error is constant (σ^2) which is not exact since it is a function of p_{ijk}. In order to remedy this situation, the arc sine transformation was applied to the square root of the proportions, that is:

$$z_{ijk} = \arcsin(\sqrt{p_{ijk}}),$$

where arcsin stands for arc sine. We have that the asymptotic variance of z_{ijk} is

$$V(z_{ijk}) \doteq \frac{1}{4\, n_{ijk}} .$$

Thus, the hypothesis of additivity of main effects to be tested becomes

$$H_o: \begin{cases} z_{ijk} = \Theta_i + \phi_j + e_{ijk} \\ \{e_{ijk}\} \text{ are independent } N(0, \frac{1}{4\, n_{ijk}}) \end{cases}$$

and minimizing

$$\sum_{ijk} n_{ijk} (z_{ijk} - \Theta_i - \phi_j)^2$$

with respect to Θ_i and ϕ_j we find the same expressions as in equation (3.3) and (3.4) but with p_{ijk} replaced by z_{ijk} and α_i replaced by Θ_i in equation (3.3). The statistic for testing additivity of effects is, in this case,

$$\chi^2 = 4\{ \sum_{ij} n_{ij.} (\sum_k \frac{n_{ijk}\, z_{ijk}}{n_{ij.}} - \sum_{ik} \frac{n_{ijk}\, z_{ijk}}{n_{.j.}})^2 - \sum_{ij} n_{ij.} (\hat{\Theta}_i - \sum_i \frac{n_{ij.}\, \hat{\Theta}_i}{n_{.j.}})^2\}$$

which has a Chi Square distribution with (I-1)(J-1) degrees of freedom.

The three main contrasts to be tested are of the same form as in equation (3.5), (3.6) and (3.7) but with α_i replaced by Θ_i, that is,

$$\psi_1 = \Theta_1 + \Theta_2 - \Theta_3 - \Theta_4,$$

$$\psi_2 = \Theta_1 - \Theta_2 + \Theta_3 - \Theta_4,$$

$$\psi_3 = \Theta_1 - \Theta_2 - \Theta_3 + \Theta_4.$$

Thus, the statistics for testing $\psi_\ell = 0$ ($\ell = 1, 3$) are

$$\chi^2_\ell = \frac{4\hat{\psi}^2_\ell}{\underline{h}'_\ell \, C^{-1} \, \underline{h}_\ell}, \qquad \ell = 1, 3$$

where $$\hat{\psi}^2_\ell = \underline{h}'_\ell \; C^{-1} \; \underline{Q} \qquad \ell = 1, 3$$

with $\underline{Q}$ defined as in (3.9) but with p_{ijk} replaced by z_{ijk}; $\underline{h}'_\ell$ and C being defined as in equation (3.8) and (3.10) respectively. The statistics χ^2_ℓ ($\ell = 1, 3$) have Chi Square distributions with one degree of freedom.

4. COMPARISON OF THE TWO MODELS

The following tables give the two statistics along with their critical point at the 5% level of error. The sets of data on which the ANOVA models were used are as follows:

i) Non-response rate to the Labour Force questionnaire; all categories of non-response are included here, that is TA's, N1's to N5's (LF N-R, in the tables), see [3] for description of codes.

ii) Refusal rate to the Labour Force questionnaire, that is, N2's (LF refusal, in the tables).

iii) Error rates on the Labour Force questionnaire as detected by the visual edit (LF errors, in the tables).

iv) Response rate to the Consumers Finance questionnaire (CF resp. in the tables).

v) Response rate to the Job Mobility questionnaire (JM resp., in the tables).

Furthermore, the tests were conducted for four consecutive months on the Labour Force data: April to July.

Four tables are presented which correspond to the four tests as described in section 3, that is, tests of the hypothesis H_o, H_1, H_2 and H_3. The asterisk on the right side of the tables indicates which tests are in disagreement.

Table 4: Test of Additivity of Main Effects

Month of Test	Type of Data \ Type of Variable	Using p as variable Statistic	Using p as variable Critical Point	Using z as variable Statistic	Using z as variable Critical Point
April	LF N-R LF refusal LF errors	0.70 0.42 1.89	2.28	6.23 14.94 14.69	21.03
May	LF N-R LF refusal LF errors	1.44 0.65 1.94	2.28	16.75 23.06 24.21	21.03 * * *
June	LF N-R LF refusal LF errors	2.55 1.55 1.64	2.54	23.69 22.23 15.48	16.92 *
July	LF N-R LF refusal LF errors	0.53 0.95 1.97	3.00	1.68 14.16 8.55	12.59 *
April	CF resp.	2.34	2.28	56.87	21.03
July	JM resp.	2.31	3.00	34.98	12.59 *

From Table 4, we find that seven sets of data using z as variable cannot be considered to have additive effects (May LF refusal and LF errors, June LF N-R and LF refusal, July LF refusal, April CF resp. and July JM resp.) while, if we use p as variable, only two sets of data (June LF N-R and April CF resp.) do not have additive effects. Five sets of data, therefore, have disagreeing tests.

Table 5: Test of Interaction Between Months and Procedures

Month of Test	Type of Data \ Type of Variable	Using p as variable Statistic	Using p as variable Critical Point	Using z as variable Statistic	Using z as variable Critical Point
April	LF N-R	0.93	4.35	0.85	3.84
	LF refusal	0.61		1.13	
	LF errors	3.85		2.17	
May	LF N-R	0.12	4.35	0.10	3.84 *
	LF refusal	3.05		4.86	
	LF errors	9.01		10.81	
June	LF N-R	2.79	4.49	2.29	3.84
	LF refusal	0.08		0.01	
	LF errors	0.33		0.15	
July	LF N-R	1.97	4.75	0.83	3.84
	LF refusal	0.00		0.32	
	LF errors	0.31		0.56	
April	CF resp.	0.58	4.35	0.97	3.84
July	JM resp.	0.12	4.75	0.00	3.84

We find in Table 5 that May LF refusal show an interaction between months and procedures when using z as variable and May LF errors have interaction whether p or z is used as variable. All other sets of data indicate no interaction with either variable. Referring to Table 4 we find that May LF refusal and LF errors do not conform to an additive model when using variable z.

Table 6: Test of Regular Procedure vs Special Procedure

Month of Test	Type of Data \ Type of Variable	Using p as variable Statistic	Using p as variable Critical Point	Using z as variable Statistic	Using z as variable Critical Point
April	LF N-R LF refusal LF errors	4.73 0.01 0.98	4.35	4.30 0.76 0.39	3.84
May	LF N-R LF refusal LF errors	0.77 0.05 0.50	4.35	0.58 0.87 0.91	3.84
June	LF N-R LF refusal LF errors	0.80 1.15 16.31	4.49	0.60 0.00 17.94	3.84
July	LF N-R LF refusal LF errors	8.68 0.63 1.49	4.75	4.47 0.04 1.29	3.84
April	CF resp.	0.02	4.35	0.08	3.84
July	JM resp.	0.22	4.75	0.20	3.84

Table 6 indicates that three sets of data (April LF N-R, June LF errors and July LF N-R) show significant differences between regular and special procedure, and this, for both p and z variables. The latter three sets of data have been found, in Table 4, to conform to the additive model and, in Table 5, to have no interaction between months and procedures.

Table 7: Test of One Month vs Two Consecutive Months

Month of Test	Type of Data \ Type of Variable	Using p as variable Statistic	Using p as variable Critical Point	Using z as variable Statistic	Using z as variable Critical Point
April	LF N-R	0.42		0.20	
	LF refusal	2.56	4.35	3.89	3.84 *
	LF errors	5.82		3.48	*
May	LF N-R	0.60		0.55	
	LF refusal	1.00	4.35	1.90	3.84
	LF errors	3.37		3.94	*
June	LF N-R	1.44		0.91	
	LF refusal	1.07	4.49	0.76	3.84
	LF errors	4.69		5.39	
July	LF N-R	27.69		14.98	
	LF refusal	0.49	4.75	0.17	3.84
	LF errors	8.42		7.17	
April	CF resp.	0.46	4.35	0.86	3.84
July	JM resp.	6.27	4.75	13.42	3.84

Table 7 reveals that seven sets of data have statistics significantly different from zero with at least one variable (p or z). Actually, three of them (April LF refusal and LF errors and May LF errors) have significant tests with one of the two variables and the four others (June LF errors, July LF N-R and LF errors and July JM resp.) have both tests significantly different from zero. For the latter seven sets of data, we found from Table 4, that two of them (May LF errors and July JM resp.) do not satisfy the assumption of additivity of main effects when using z as variable.

5. CONCLUSION

If we consider that the assumptions of additivity of main effects and zero interaction between months and procedures must be satisfied in order to test for main effects, the model using p as variable seems to better suit the purpose of the analysis. But, one must take the results found with p as variable with caution since the error variance is actually a function of p. On the other hand, though the transformation yielding variable z has the property of stabilizing the error variance, it is difficult to give a meaning to the resulting effects.

Many computations and tests have been made on the sets of data in order to find trends that could provide explanations for the disagreement of the two tests but in all cases the results have been negative. We easily see, though, that the test yielding the most disagreement is the test of additivity of main effects. Actually, this situation is quite normal since, due to the type of transformation, the effects considered when p is the variable are completely different from those considered when z is the variable. The fact that no particular trends has been found between disagreeing tests is may be due to the low sensitivity of the tests since the number of degrees of freedom is fairly small in all cases.

6. ACKNOWLEDGMENT

The author is grateful to the referee and the editors for very useful and constructive comments.

RESUME

L'analyse des données du Projet d'Expérimentation Méthodologique (PEM) Phase II fut faite sous forme d'analyse de variance. Mais puisque les données sont des proportions et qu'un modèle d'analyse de variance utilisant ces données ne fournit que des résultats approximatifs, deux modèles furent utilisés pour l'analyse. L'analyse de la variance fut donc faite utilisant les données sous forme de proportions ainsi qu'en les transformant au moyen de la transformation arc sinus de la racine carrée. Le présent article détaille les deux modèles utilisés et les compare empiriquement au moyen des données du PEM Phase II.

REFERENCES

[1] Cochran, W.G., "The Analysis of Variance When Experimental Errors Follow the Poisson or Binomial Laws", Annals of Mathematical Statistics, September 1940, Vol. 11, pp. 335-347.

[2] Dyke, G.V. and Patterson, H.D., "Analysis of Factorial Arrangements when the Data are Proportions", Biometrics, March 1952, vol. 8, pp. 1-12.

[3] Sugavanam, R., "Non-Interview Patterns in the Canadian Labour Force Survey", Survey Methodology Journal, June 1975, vol. 1, pp. 44-59.

[4] Tessier, R., Tremblay, V., and Maguire, E., "Test on Respondent Burden, Methods Test Panel Phase II, Methodology", Household Surveys Development Staff, Statistics Canada.

[5] Tessier, R., Burrows, D., and Tremblay, V., "Test on Respondent Burden, Methods Test Panel Phase II, Analysis", Household Surveys Development Staff, Statistics Canada.

ESTIMATION OF PROCESS AVERAGE IN ATTRIBUTE SAMPLING PLANS

P.D. Ghangurde
Household Surveys Development Division

Exact formulae for bias and mean square error of an estimator of process average in single sampling with rectification for finite lots are obtained. Efficiency of the estimator as compared to an unbiased estimator based on the first sample is obtained for a number of values of lot size, sample size, acceptance number and process average used in sampling plans in quality control of data processing.

1. INTRODUCTION

In single sampling with rectification plans a sample of fixed size is drawn from a lot. If the number of defectives in the sample is less than or equal to c, an acceptance number, the lot is accepted; otherwise the lot is completely verified and rectified. The sample size and acceptance number are determined to minimize average amount of inspection for a given lot size and proportion of defectives [3]. These plans are appropriate in situations where inspection is nondestructive and rectification is not costly.

Maximum likelihood estimators (m.l.e.) of process average for single and double attribute sampling plans have been given in the literature assuming constant process average and large lot sizes (see e.g. [1] and [4]). Formulae for asymptotic variance of the m.l.e. are also obtained under these assumptions. However, there are many industrial processes in which process average can change even when the process is in control [5] and sampling plans have to be altered according to changes in the process average.

The estimator proposed in this paper is appropriate in situations in which fraction defective and lot size could vary considerably from lot to lot. Exact expressions for bias and variance of the estimator can be easily derived.

2. ESTIMATOR OF FRACTION DEFECTIVE

Consider a lot of size N and proportion of defectives P from which a random sample of size n_1 is drawn without replacement. If the number of defectives, $x_1 \leq c$, the acceptance number, the lot is accepted. If $x_1 > c$ the lot is completely verified and hence P can be determined without any sampling error. An estimator of P can be defined as

$$e_1 = \frac{x_1}{n_1} \alpha + P(1 - \alpha), \tag{2.1}$$

where α is a random variable defined by

$$\alpha = \begin{cases} 1 & \text{if } x_1 \leq c, \\ 0 & \text{if } x_1 > c. \end{cases} \tag{2.2}$$

The estimator e_1 is of the same form as m.l.e. in double sampling scheme discussed in section 4 (where N is assumed infinite) when n_2, the second sample size, tends to infinity.

3. BIAS AND MEAN SQUARE ERROR

$$E(e_1) = \frac{1}{n_1} E(x_1 | x_1 \leq c)\phi + (1 - \phi)P,$$

where

$$\phi = P[x_1 \leq c]. \tag{3.1}$$

Hence,

$$\begin{aligned} \text{Bias } (e_1) &= E(e_1) - P \\ &= \phi\left[\frac{E(x_1 | x_1 \leq c)}{n_1} - P\right]. \end{aligned} \tag{3.2}$$

Let L(> 1) lots with known sizes, N_i, be inspected by a single sampling with rectification plan (n_1, c). The process average can be defined as

$$\bar{P} = \sum_{i=1}^{L} \pi_i P_i ,$$

where

π_i = known proportion of ith lot size in the total items in L lots,

P_i = fraction defective of ith lot,

and $\sum_{i=1}^{L} \pi_i = 1.$

The estimator e_1 is defined as

$$e_1 = \sum_{i=1}^{L} \pi_i \left[\frac{x_{1i}\,\alpha_i + X_{1i}(1 - \alpha_i)}{n_1\,\alpha_i + N_i(1 - \alpha_i)} \right], \tag{3.3}$$

where

$$\alpha_i = \begin{cases} 1 & \text{if } x_{1i} \leq c, \\ 0 & \text{otherwise} \end{cases}$$

x_{1i} = defectives in sample of size n_1 from ith lot,

X_{1i} = total number of defectives in ith lot,

N_i = size of ith lot.

Hence,

$$\text{Bias } (e_1) = \sum_{i=1}^{L} \pi_i \phi_i \left[\frac{E(x_{1i} | x_{1i} \leq c)}{n_1} - P_i \right] \tag{3.4}$$

where

$$\phi_i = P[x_{1i} \leq c].$$

In order to study behavior of Bias (e_1) for single lot we consider Poisson approximation.

Let $\lambda = n_1 P$ then $\phi = e^{-\lambda} \sum_{x_1=0}^{c} \lambda^{x_1}/x_1!$ and

$$E(x_1 | x_1 \leq c) = \frac{e^{-\lambda}}{\phi} \sum_{x_1=1}^{c} \lambda^{x_1}/(x_1-1)! \ .$$

The bias expression in (3.2) takes the form Bias $(e_1) = - e^{-\lambda} \lambda^{c+1}/n_1 c!$. Relative Bias (e_1) can be defined as Bias $(e_1)/P$ and is given by Relative Bias $(e_1) = - e^{-\lambda} \lambda^c/c!$. Thus Bias (e_1) is negative and for a given optimum sampling plan (n_1, c) obtained for certain process average P the absolute value of Relative Bias (e_1) is a monotonic increasing function of P for $P < \frac{c}{n_1}$ and a monotonic decreasing function of P for $P > \frac{c}{n_1}$. Though a plan (n_1, c) is determined to give minimum inspection for certain process average P, the fraction defective of a lot could be different from P and hence study of behavior of bias as a function of P for a given plan is of practical importance.

Table 1 gives numerical values of Bias (e_1) and Relative Bias (e_1) obtained by using binomial probabilities in formula (3.2) for single lot for various values of P, N and plans (n_1, c) used in quality control of data capture by keypunch and key-edit. The values of n_1 and c are the optimum ones giving minimum average inspection for given lot size N and fraction defective P in single sampling with rectification assuring 3% AOQL [3]. The binomial probabilities tabulated in [6] are used in the calculations. It can be seen that absolute value of Relative Bias (e_1) decreases as n_1 and c are increased for a given P.

The tabulated values of Bias (e_1) for various plans show the extent of bias of e_1 for various plans and values of P. The bias of the estimator in (3.3) can be estimated from rejected lots by

$$\widehat{\text{Bias}}\ (e_1) = \sum_{i=1}^{L} \pi_i (1 - \alpha_i)(P_i - \frac{x_{1i}}{n_{1i}}). \tag{3.5}$$

For single lot

$$\begin{aligned}
V(e_1) &= E[e_1 - E(e_1)]^2 \\
&= E[\frac{x_1}{n_1}\alpha + P(1-\alpha) - \frac{1}{n_1} E(x_1|x_1 \le c)\phi - P(1-\phi)]^2 \\
&= E[\frac{1}{n_1}(x_1\alpha - E(x_1|x_1 \le c)\phi) - P(\alpha - \phi)]^2 \\
&= \frac{1}{n_1^2} E[x_1\alpha - E(x_1|x_1 \le c)\phi]^2 + P^2 E(\alpha - \phi)^2 \\
&\quad - \frac{2P}{n_1} . E[(x_1\alpha - E(x_1|x_1 \le c)\phi)(\alpha - \phi)] \\
&= \frac{\phi}{n_1^2} [E(x_1^2|x_1 \le c) - E^2(x_1|x_1 \le c)\phi] + P^2\phi(1-\phi) \\
&\quad - \frac{2P}{n_1} . E(x_1|x_1 \le c)\phi(1-\phi) \qquad (3.6) \\
&= \frac{\phi}{n_1^2} [E(x_1^2|x_1 \le c) - E^2(x_1|x_1 \le c)\phi] - \phi(1-\phi)P^2 \\
&\quad - 2P(1-\phi)[\text{Bias } (e_1)]
\end{aligned}$$

The mean square error of e_1 is given by

$$\text{MSE } (e_1) = V(e_1) + [\text{Bias } (e_1)]^2 \tag{3.7}$$

This is the expression used in numerical efficiency comparison of e_1 with e_1' (see Table 1) which is based on first sample only and given by $e_1' = x_1/n_1$. Its variance for finite lot sizes is given by

$$V(e_1') = \frac{N-n_1}{N-1} \cdot \frac{P(1-P)}{n_1} . \tag{3.8}$$

Though finite population correction is used in $V(e_1')$, the MSE (e_1) is obtained by using binomial probabilities from Tables in [6]. Since $N/n_1 > 10$ for most entries in Table 1, the binomial approximation to hypergeometric probabilities is expected to be very close.

4. ESTIMATION IN DOUBLE SAMPLING

Consider a lot of size N from which a random sample of size n_1 is drawn without replacement. If the number of defectives, $x_1 \leq c$, the acceptance number, the lot is accepted. If $x_1 > c$ a second sample of size n_2 is drawn and the number of defectives, x_2, is observed. In practice, in double sampling, if $c < x_1 \leq c_1$, where c_1 is another acceptance number, a second sample of size n_2 is drawn. The lot is accepted if $x_1 + x_2 \leq c_1$ and rejected if $x_1 + x_2 > c_1$. The above double sampling scheme is considered for simplicity, since the purpose is to obtain the results of single sampling with rectification for large lots as a limiting case of double sampling when $n_2 \to \infty$.

The likelihood of the sample, assuming large N, is given by

$$L(n_1, n_2, x_1, x_2, P) = \begin{cases} \binom{n_1}{x_1} P^{x_1} (1-P)^{n_1-x_1}, & \text{if } x_1 \leq c \\ \binom{n_1}{x_1}\binom{n_2}{x_2} P^{x_1+x_2} (1-P)^{n_1+n_2-x_1-x_2} & \text{if } x_1 > c \end{cases}$$

Let α be a random variable defined as in (2.2). Hence

$$L(n_1, n_2, x_1, x_2, P) = \left[\binom{n_1}{x_1} P^{x_1} (1-P)^{n_1-x_1}\right]\left[\binom{n_2}{x_2} P^{x_2} (1-P)^{n_2-x_2}\right]^{(1-\alpha)} .$$

Differentiating L with respect to P the m.l.e. is obtained as

$$e_2 = \frac{x_1 + x_2(1 - \alpha)}{n_1 + n_2(1 - \alpha)} \tag{4.1}$$

$$\begin{aligned} E(e_2) &= E[\frac{x_1}{n_1} |\alpha=1]\phi + E[\frac{x_1 + x_2}{n_1 + n_2} |\alpha=0](1 - \phi) \\ &= \frac{1}{n_1} E(x_1|x_1 \leq c)\phi + \frac{1}{(n_1+n_2)} [E(x_1|x_1 > c)(1 - \phi) + n_2 P(1 - \phi)] \\ &= \frac{1}{n_1(n_1+n_2)} [n_1^2 P + n_2 E(x_1|x_1 \leq c)\phi + n_1 n_2 P(1 - \phi)] \end{aligned}$$

since $\quad n_1 P = E(x_1|x_1 \leq c)\phi + E(x_1|x_1 > c)(1 - \phi).$

Hence,

$$\begin{aligned} \text{Bias}(e_2) &= E(e_2) - P \\ &= (\frac{n_2}{n_1 + n_2})\phi [\frac{E(x_1|x_1 \leq c)}{n_1} - P]. \end{aligned} \tag{4.2}$$

As $n_2 \to \infty$ (4.2) takes the same form as (3.2).

For L lots assuming the same plan (n_1, c) with the same notations as before

$$e_2 = \sum_{i=1}^{L} \pi_i [\frac{x_{1i} + x_{2i}(1 - \alpha_i)}{n_1 + n_2(1 - \alpha_i)}] \tag{4.3}$$

with obvious meanings for x_{1i}, x_{2i}, α_i, $i = 1, 2, \ldots, L$. Assuming fraction defective for each of L lots to be P, the m.l.e. of P for double sampling can be obtained as

$$e_3 = \frac{\sum_{i=1}^{L} [x_{1i} + x_{2i}(1 - \alpha_i)]}{\sum_{i=1}^{L} [n_1 + n_2(1 - \alpha_i)]} . \tag{4.4}$$

This is the estimate of P generally used in the literature. If the process average of individual lots is known to be different (4.3) is more appropriate than (4.4). The estimator (4.3) takes the form (3.3) for single sampling with rectification. The expression for Bias (e_3) for general L is complicated. For simplicity we consider the case L = 2 and for single sampling with rectification obtain for large lot sizes

$$\text{Bias } (e_3) = \phi_1 \phi_2 \left[\frac{E(x_{11}|x_{11} \leq c) + E(x_{12}|x_{12} \leq c)}{2n_1} - \pi_1 P_2 - \pi_2 P_1 \right]$$

$$+ (P_2 - P_1)(\phi_1 \pi_1 - \phi_2 \pi_2). \qquad (4.5)$$

For L = 2 and single sampling with rectification Bias (e_2) is obtained as

$$\text{Bias } (e_2) = \pi_1 \phi_1 \left[\frac{E(x_{11}|x_{11} \leq c)}{n_1} - P_1 \right] + \pi_2 \phi_2 \left[\frac{E(x_{12}|x_{12} \leq c)}{n_1} - P \right].$$

When $P_1 = P_2$, $\pi_1 = \pi_2 = 1/2$ absolute value of Bias (e_3) is less than that of Bias (e_2). Since for given (n_1, c), ϕ decreases as P increases the contribution of second term in (4.5) is positive when $P_1 \neq P_2$ and $\pi_1 = \pi_2 = 1/2$. It seems that absolute value of Bias (e_3) would be lesser than that of Bias (e_2). However, no conclusions can be drawn for the general case of L > 2.

We now obtain variance of e_2 for single lot,

$$V(e_2) = V\, E(e_2|\alpha) + E\, V(e_2|\alpha)$$

After some algebra and reduction we obtain

$$V(e_2) = \frac{\phi}{n_1^2}\,[E(x_1^2|x_1 \le c) - E^2(x_1|x_1 \le c)\phi]$$

$$+ \frac{(1-\phi)}{(n_1+n_2)^2}\,[E(x_1^2|x_1 > c) - E^2(x_1|x_1 > c)(1-\phi)]$$

$$+ \frac{(1-\phi)\,n_2P[1 - P + n_2P\,\phi]}{(n_1+n_2)^2}$$

$$+ \frac{2n_2P\,\phi(1-\phi)\,E(x_1|x_1 > c)}{(n_1+n_2)^2}$$

$$- \frac{2\,E(x_1|x_1 \le c)\,n_2P\,\phi(1-\phi)}{n_1(n_1+n_2)}$$

$$- \frac{2\,E(x_1|x_1 \le c)\,E(x_1|x_1 > c)\phi(1-\phi)}{n_1(n_1+n_2)} \qquad (4.6)$$

As $n_2 \to \infty$ $V(e_2)$ takes the limiting form

$$V(e_2) = \frac{\phi}{n_1^2}\,[E(x_1^2|x_1 \le c) - E^2(x_1|x_1 \le c)\phi]$$

$$+ \phi(1-\phi)P^2 - \frac{2PE(x_1|x_1 \le c)\phi(1-\phi)}{n_1}$$

which is the same as (3.6)

5. ACKNOWLEDGMENTS

The author would like to acknowledge the referee for some helpful comments.

Table 1:

c = 1

P	B	n	ϕ	Bias(e_1)	Relative Bias(e_1)	MSE(e_1)	V(e_1')	Eff(e_1)
.004	900	26	.995122	-.000362	-.090466	.000126	.000149	118.089
.010	775	26	.972277	-.001945	-.194455	.000241	.000368	153.003
.012	600	26	.961318	-.002662	-.221843	.000267	.000437	163.808
.014	496	26	.948979	-.003444	-.246032	.000289	.000504	174.459
.016	424	26	.935418	-.004276	-.267262	.000309	.000570	184.658
.018	361	25	.926016	-.005028	-.279357	.000347	.000660	190.400
.020	310	25	.911355	-.005911	-.295575	.000365	.000723	198.360
.022	277	25	.895892	-.006811	-.309575	.000382	.000786	205.699
.024	250	25	.879735	-.007717	-.321527	.000400	.000847	211.900
.026	230	24	.871832	-.008483	-.326255	.000443	.000949	214.370
.028	230	24	.855512	-.009384	-.335127	.000462	.001020	220.862

c = 2

P	N	n	ϕ	Bias(e_1)	Relative Bias(e_1)	MSE(e_1)	V(e_1')	Eff(e_1)
.010	900	42	.991416	-.000549	-.054856	.000200	.000225	112.412
.012	900	42	.985999	-.000874	-.072854	.000227	.000269	118.879
.014	900	42	.979009	-.001280	-.091442	.000249	.000314	126.140
.016	900	42	.970414	-.001762	-.110118	.000267	.000358	134.104
.018	900	42	.960216	-.002313	-.128475	.000282	.000402	142.662
.020	750	42	.948450	-.002924	-.146190	.000294	.000441	150.227
.022	637	42	.935178	-.003586	-.163010	.000304	.000479	157.810
.024	563	42	.920478	-.004290	-.178743	.000312	.000517	165.526
.026	500	41	.909603	-.004907	-.188730	.000333	.000568	170.374
.028	449	41	.893128	-.005657	-.202020	.000341	.000605	177.172
.030	409	41	.875552	-.006420	-.214010	.000349	.000640	183.453

RESUME

On définit les formules exactes pour calculer le biais et l'erreur quadratique moyenne d'un estimateur de moyenne du processus dans un échantillonnage unique avec correction pour les lots finis. L'efficacité de l'estimateur par rapport à un estimateur non biaisé fondé sur le premier échantillon s'obtient pour un certain nombre de valeurs de la taille des lots de l'échantillon, du nombre d'acceptation et de la moyenne du processus utilisées dans les schémas d'échantillonnage servant au contrôle qualitatif du traitement des données.

REFERENCES

[1] Cohen, A.C. (1970), "Curtailed Attribute Sampling", Technometrics, 12, 295-298.

[2] Duncan, A.J. (1965), "Quality Control and Industrial Statistics", R.D. Irwin.

[3] Dodge and Romig, (1959), "Sampling Inspection Tables, Single and Double Sampling", Second Edition, Page 60, John Wiley and Sons, Inc., New York.

[4] Phatak, A.G. and Bhatt, M.M., (1967), "Estimation of Fraction Defective in Curtailed Sampling Plans", Technometrics, 9, 219-288.

[5] Wetherill, G.B., (1969), "Sampling Inspection and Quality Control", Page 60, Methuen and Co. Ltd., London.

[6] Weintraub, S., (1963), "Tables of the Cumulative Binomial Probability Distribution for Small Values of P", Free Press of Glencoe, New York.

LIST OF REFEREES

The Editorial Board wish to thank the following persons who have served as referees during the past year.

Cairns, M.
Chauvin, G.
Drew, D.
Ghangurde, P.D.
Gough, H.
Gray, G.B.
Hidiroglou, M.A.
Satin, A.
Singh, M.P.
Sugavanam, R.
Tessier, R.
Timmons, P.F.
Tremblay, V.

DATE DUE
DATE DE RETOUR

DEC 31 1985			
JUL 27 1988			
JAN 4 1989			

SURVEY METHODOLOGY

June 1975 Volume 1 Number 1

A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.

C O N T E N T S

SURVEY METHODOLOGY

December 1975 Volume 1 Number 2

A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.

CONTENTS